Notes on Bartik Instruments*

Paul Goldsmith-Pinkham Isaac Sorkin Henry Swift

This version: January 10, 2017.

Preliminary.

Abstract

We study the econometric properties of the Bartik instrument, a commonly used approach for identification in economics. We develop a formal econometric framework for the Bartik instrument and show an exact relationship between the Bartik approach and an overidentified GMM estimator. We also use our structure to highlight several points that are relevant for practice. With these insights, we develop a checklist of recommendations for how to implement the Bartik instrument and how to test the plausibility of the exclusion restriction.

*Goldsmith-Pinkham: Federal Reserve Bank of New York. Email: paulgp@gmail.com. Sorkin: Department of Economics, Stanford University. Email: sorkin@stanford.edu. Swift: Unaffiliated. Email: henryswift@gmail.com. The views expressed are those of the authors and do not necessarily reflect those of the Federal Reserve Bank of New York or the Federal Reserve Board. All errors are our own.

The Bartik, or shift-share, instrument is named after Bartik (1991) and popularized in Blanchard and Katz (1992).[1] These papers defined the Bartik instrument as the predicted local employment growth created by the interaction of local industry employment shares with national industry employment growth rates. This instrument and variants have since been widely used across many fields in economics, including labor, public, macroeconomics, international trade, and finance. (See Table 1.)

Despite the ubiquity of the instrument in practice, there is little work laying out exactly what the Bartik instrument is, what the identifying assumptions are, and why it works. The goal of this paper is to fill this gap. We first develop an econometric structure that defines what we mean by a Bartik instrument, and show that this set-up encompasses several instruments used in applied economics. Second, we use our structure to specify the identifying assumption in the Bartik instrument. Third, we show why Bartik is effective. Finally, our framework is informative for a number of smaller practical issues.

We begin by developing a formal econometric structure for the Bartik instrument. To fix ideas, let $x_l = Z_l' G_l = \sum_k z_{lk} g_{lk}$, where x_l is the location l employment growth rate, Z_l is the vector of location-specific industry shares (where z_{lk} is the k^{th} entry in Z_l) and G_l is the vector of location-specific growth rates (where g_{lk} is the k^{th} entry). The Bartik instrument is constructed by replacing the location-specific G_l with some national average, $\overline{G}$, so that $B_l = Z_l' \overline{G}$, and the researcher uses B_l as an instrument for x_l.

We define a Bartik-like instrument as an instrument in which, as in the previous paragraph, the researcher exploits the inner product structure of the endogenous variable to construct an instrument. This encompasses several instruments in the applied literature, which are not always labelled as Bartik instruments *per se*. First, the simulated instrument of Currie and Gruber (1996b) interacts local laws with national population characteristics. Second, the "immigrant enclave" instrument introduced by Altonji and Card (1991) interacts initial immigrant composition of a place with immigration flows from origin countries. Third, the "China shock" instrument of Autor, Dorn, and Hanson (2013) interacts local industry composition with the growth of Chinese imports to European countries. Fourth, researchers, such as Greenstone, Mas, and Nguyen (2015), have interacted pre-existing bank lending shares with changes in bank lending volumes to instrument for credit supply. Fifth, Acemoglu and Linn (2004) interact age-group spending patterns with demographic changes to instrument for market size. Finally, the "judge" instrument interacts a case's assigned judge with a measure of the leniency of the judge. While many substantive questions have been studied using Bartik-like instruments, for concreteness our running example is the canonical application of national industry growth rates interacted with local

[1] Arguably the Bartik instrument pre-dates Bartik (1991) and is also sometimes known as a shift-share instrument. See Freeman (1980) for an earlier use.

industry composition.

We first show that the exogeneity of the Bartik instrument relies on the exogeneity of the local industry shares, while the relevance of the Bartik instrument relies on the variation in the national industry growth rates. As an extreme example, with only two industries, the Bartik instrument is numerically equivalent to using either industry share as an instrument. Moreover, for an arbitrary number of industries, the Bartik instrument in two stage least squares is numerically equivalent to a generalized methods of moments estimator with the industry shares as instruments, where the weight matrix is constructed from the national industry growth weights. This finding sheds light on a common argument that the national nature of the average industry growth rates avoids correlation with local economic conditions.[2] Instead, we show that researchers do not need to make arguments about the properties of the national industry growth rates to support the validity of the Bartik instrument, and should instead focus on arguing for the the exogeneity of the industry shares.

This set-up clarifies the endogeneity concerns that Bartik can address. Consder the local-industry growth rates as the sum of two terms: the national industry growth rate and an idiosyncratic industry-location specific growth rate. The Bartik instrument addresses the concern that the idiosyncratic industry-location specific growth rates are correlated with the error term. For example, suppose there are two industries, service and manufacturing, and that both national growth rates are zero. Then, if a location has positive growth in both manufacturing and services, then we might be concerned that this reflects a supply shock in the location. If this supply shock is uncorrelated with the industry shares, then using the Bartik instrument can address this issue.

We then show that the Bartik instrument should be thought of as an estimator designed to overcome the problem of a high-dimensional first-stage, rather than an instrument. Given that the Bartik instrument assumes the exogeneity of industry shares, the question is why the researcher constructs the Bartik instrument, rather than using the industry shares as instruments directly. Our answer is that it reduces the dimension of the first stage. Under the assumption of the exogeneity of industry shares (Z), the first-stage estimation problem is to estimate $\mathbb{E}_l[x|Z]$, where this expectation is taken over all locations (subscripted by l). With the inner product structure of x, this first-stage estimation problem can be rewritten as follows:

$$\mathbb{E}_l[x|Z] = \mathbb{E}_l[Z'G|Z] = Z'\mathbb{E}_l[G|Z].$$

[2]For example, Bound and Holzer (2000, pg. 31): "The [Bartik] index should capture exogenous shifts in local labor demand that are predicted by the city-specific industry mix, while avoiding the endogeneity associated with local employment growth rates. We use this index as an instrument for the overall local employment growth."

The dimension of this first-stage, however, is proportional to the number of industries, and thus potentially large. The Bartik instrument avoids a high-dimensional first-stage by focusing on the $\mathbb{E}_l[G|Z]$ component and estimating it in the following way:

$$\hat{\mathbb{E}}_l[G|Z] = \mathbb{E}_l[G].$$

This approximation ignores the information in how growth rates vary by industry composition; for example, it might be that the restaurant industry grows quickly in locations with a large entertainment industry presence, or the restaurant industry growth rate depends on the initial restaurant industry share. By ignoring this information, the Bartik instrument avoids the appearance of a high-dimensional first stage.[3]

Nevertheless, Bartik is still leveraging a high-dimensional set of instruments and so thinking about the implicit dimensionality of the first stage provides guidance as to how finely to divide industries and locations. In finite samples, having a high-dimensional first stage can induce bias by overfitting the first stage (i.e., Angrist, Imbens, and Krueger (1999)). In practice, researchers should make choices to avoid having "too many" parameters to estimate relative to the amount of data. In finite, for example, by using a very fine set of industry shares with very few firms in each industry, in finite samples the Bartik estimator may be a poor approximation. A similar observation explains the value of the common practice of using a leave-one-out estimator to construct the national averages.

Given that the identifying assumption is the exogeneity of industry shares, our structure shows that if changes in local industry composition are partially driven by endogenous shocks, then updating industry shares may bias estimates. For example, if the endogeneous component of growth is serially correlated, then the industry shares are mechanically correlated with this endogeneous component, rendering the instrument invalid. Hence, researchers should use the earliest potential version of their industry shares to avoid this correlation.

By clarifying that the identifying assumption is the exogeneity of industry shares, our structure provides guidance to researchers about how to explore the plausibility of their research design. We recommend two tests in particular. First, researchers should present balance tests in terms of industry shares to show that the industry shares are not related to other location characteristics. Since the Bartik instrument is typically in changes, these tests should examine both levels and changes. Second, researchers should examine pre-trends. Testing for pre-trends in the Bartik setting cannot be done directly, since if the instrument is

[3]Given the same identifying assumption as the Bartik instrument, there are potentially more powerful estimators that exploit knowledge of the industry growth rates across different locations. However, we show that in this context, industry-specific growth rates explain almost 25% percent of the variance of industry-by-location growth rates (using PUMAs and 3 digit industries). As a result, the use of industry averages in Bartik is a surprisingly well-supported approximation.

valid and correlated through time, then parallel trends will not hold. We propose a simple method for partialling out the effect of previous values of the instrument such that under the assumed exogeneity conditions, parallel trends ought to hold.

Finally, we illustrate our points about the Bartik instrument in the empirical context of using the Bartik instrument as an instrument to estimate labor supply and in simulations. In the empirical example, we show that 1) the industry shares are correlated with many observable characteristics (including education), 2) controlling for these observable differences attenuates estimates, and 3) there appear to be pre-trends. Via simulation, we show that 1) endogeneity of growth rates is not necessarily a problem for identification, 2) a small variance of the industry common component leads to noisy estimates, and 3) a small number of locations relative to the number of industries also appears to lead to bias.

We note three limitations to our analysis. First, we study a setting where locations are independent. This focus means that we ignore the possibility of spatial spillovers and spatial correlation. Second, we consider settings where the number of locations grows large, but the number of industries is fixed. There are potentially many instruments issues that arise when we allow both to grow large. Finally, we will not write down or try to find an economic model that would justify the Bartik instrument in the canonical application (see Kovak (2013) for one version).

Roadmap: Section 1 introduces our notation and shows that the key identifying assumption is the exogeneity of industry shares. Section 2 discusses estimation issues, and points out that the Bartik instrument is really an estimation approach. Section 3 lays out two tests for confounds, and shows that in the canonical application of Bartik there is a lack of balance, and pre-trends. Section 4 illustrates some of our identification results via simulation. Section 5 relates our structure to several other instruments in economics. Section 6 concludes by summarizing the results of the paper in a checklist of recommendations for practice.

1 Identification

We first focus on understanding the identifying assumptions in using a Bartik instrument. We begin by setting up the notation for the cross-sectional data case. We then specialize to the two-industry case where it is easiest to see the result that industry shares are the instrument. We then generalize in two directions: first, by showing how the result holds in a panel setting and second by showing how the result holds with an arbitrary number of industries and time periods.

1.1 Cross-sectional data

We start with the cross-sectional data case. We are interested in estimating β in the following equation:

$$y_l = \beta_0 + \beta x_l + \epsilon. \tag{1.1}$$

We consider $\{\epsilon_l, x_l\}_{l=1}^{L}$ to be independent and identically distributed with L going to infinity. In the canonical Bartik set-up, l indexes a location, y_l is wage growth, and x_l is employment growth. To make the problem interesting, we assume that $\mathbb{E}_l[x_l \epsilon_l] \neq 0$, so that the OLS estimator for β is biased and we want an instrument (where the l subscript indicates that the expectation is taken over locations).

The Bartik instrument depends on the inner product structure of x_l. Specifically, assume that there are K industries and let

$$x_l = Z_l' G_l = \sum_{k=1}^{K} z_{lk} g_{lk}, \tag{1.2}$$

where Z_l is a $K \times 1$ vector of industry-location shares and $\sum_{k=1}^{K} z_{lk} = 1$ and G_l is a $K \times 1$ vector industry-location growth rates where the k^{th} entry is g_{lk}. It is helpful to think of decomposing the industry-location growth rate as follows:

$$g_{lk} = g_k + \tilde{g}_{lk}, \tag{1.3}$$

where g_k is the industry component of growth, and $\tilde{g}_{lk}$ is the idiosyncratic industry-location component of growth.

We are now in a position to construct the Bartik instrument. Let

$$\hat{g}_k = \frac{1}{L} \sum_{l=1}^{L} g_{lk} \tag{1.4}$$

be the estimate of the national growth rate in industry k, and let $\hat{G} = (\hat{g}_1, \cdots, \hat{g}_K)'$ be the vector of national growth rates. Then the Bartik instrument is constructed as follows:

$$\hat{B}_l = Z_l' \hat{G}. \tag{1.5}$$

Hence, our set-up with two stage least squares (2SLS) is as follows:

$$y_l = \beta_0 + \beta x_l + \epsilon_l \tag{1.6}$$

$$x_l = \gamma_0 + \gamma \hat{B}_l + \eta_l. \tag{1.7}$$

The two assumptions for the 2SLS estimator are that the Bartik instrument is *relevant* so that $\mathbb{E}_l[\hat{B}_l x_l] \neq 0$, and that the instrument is *exogenous* so that $\mathbb{E}_l[\hat{B}_l \epsilon_l] = 0$, where the expectations are taken over locations. So far this is just a restatement of standard results. We now ask: what do assumptions about $\hat{B}_l$ imply about properties of Z_l and $\hat{G}$?

Two industries

It is easiest to see the answer to this question in the two industry case. Let $K = 2$. We might think of the industries as manufacturing and services. Then $z_{l2} = 1 - z_{l1}$ so that we can write Bartik explicitly:

$$\hat{B}_l = z_{l1}\hat{g}_1 + z_{l2}\hat{g}_2 \tag{1.8}$$

$$= z_{l1}\hat{g}_1 + (1 - z_{l1})\hat{g}_2 \tag{1.9}$$

$$= \hat{g}_2 + (\hat{g}_1 - \hat{g}_2)z_{l1}. \tag{1.10}$$

With two industries, using the Bartik instrument in 2SLS is formally identical to using z_{l1} as an instrument. While with more than two industries this exact equivalence does not hold, we show below that this equivalence is recovered in a generalized method of moments (GMM) set-up and the insights that we describe below also generalize.

The first stage in the two industry case is:

$$x_l = \gamma_0 + \gamma \hat{B}_l + \eta_l \tag{1.11}$$

$$x_l = \underbrace{\gamma_0 + \gamma \hat{g}_2}_{\text{constant}} + \underbrace{\gamma(\hat{g}_1 - \hat{g}_2)}_{\text{coefficient}} z_{l1} + \eta_1. \tag{1.12}$$

Hence, for exogeneity we need $\mathbb{E}[z_{l1}\epsilon_l] = 0$ and for relevance we $\hat{g}_1 - \hat{g}_2 \neq 0$ and $\gamma \neq 0$. That is, the shares are about exogeneity and the growth rates are about relevance (or, power).

It is helpful to see why in our set-up OLS might be inconsistent and thus what endogeneity concern Bartik can address. Consider the difference between β and $\hat{\beta}_{OLS}$:

$$\beta - \hat{\beta}_{OLS} = \frac{(g_1 - g_2)\sum_{l=1}^{L}((z_{1l} - \bar{z}_1))\epsilon_l}{\sum_{l=1}^{L}[(x_l - \bar{x})^2]} \tag{1.13}$$

$$+ \frac{\sum_{l=1}^{L}\left[(\tilde{g}_{1l} - \tilde{g}_{2l})z_{1l} - \left[\sum_{j=1}^{L}(\tilde{g}_{1j} - \tilde{g}_{2j})z_{1j}\right]\right]\epsilon_l}{\sum_{l=1}^{L}[(x_l - \bar{x})^2]}. \tag{1.14}$$

OLS is inconsistent when $corr((\tilde{g}_{1l} - \tilde{g}_{2l}), \epsilon_l) \neq 0$. Recall that $\tilde{g}$ means deviation from the national average. One way this correlation could be non-zero is if locations that have an unusually large gap between services and manufacturing growth rates have unusually large

amenity growth, where amenity growth is a factor affecting wages that does not directly affect employment growth.

The result also shows that it is mechanical that using national industry growth rates avoids correlation with the error term. The reason is that the national industry growth rates are a constant across locations and a constant has zero correlation with a random variable. This finding emphasizes that when researchers wish to make the case for Bartik the argumentation should be about the exogeneity of the industry shares, and not about the properties of the national growth rates.

1.2 Panel data

We now show that when Bartik is used in a panel, it is equivalent to allowing for time-variation in the weights on the industry shares. We also show that when there is serial correlation in the endogenous component of x, researchers should use the initial industry shares to estimate Bartik.

Maintaining the assumption of two industries, define the Bartik instrument so that it varies over time:

$$\hat{B}_{lt} = \bar{g}_{1t} z^0_{1l} + \bar{g}_{2t} z^0_{2l},$$

where now there are L locations and T time periods, indexed by l and t respectively. (We fix the z to an initial time period, for reasons that we explain below).

To see the relationship between the cross-sectional and the panel estimating equations, it is helpful to write the setup with place and time fixed effects:

$$y_{lt} = \alpha_l + \alpha_t + x_{lt}\beta + \epsilon_{lt} \tag{1.15}$$

$$x_{lt} = \tau_l + \tau_t + \hat{B}_{lt}\gamma + u_{lt}. \tag{1.16}$$

Now substitute in the Bartik instrument and rearrange the first stage:

$$x_{lt} = \tau_l + \tau_t + \left(\bar{g}_{1t} z^0_{1l} + \bar{g}_{2t} z^0_{2l}\right)\gamma + u_{lt} \tag{1.17}$$

$$= \tau_l + \underbrace{\left(\tau_t + \bar{g}_{1t}\gamma\right)}_{\tilde{\tau}_t} + z^0_{2l} \underbrace{\left(\bar{g}_{2t} - \bar{g}_{1t}\right)\gamma}_{\tilde{\gamma}_t} + u_{lt} \tag{1.18}$$

$$= \tau_l + \underbrace{\left(\tau_t + \bar{g}_{1t}\gamma\right)}_{\tilde{\tau}_t} + z^0_{l2} \underbrace{\Delta_{g,t}\gamma}_{\tilde{\gamma}_t} + u_{lt}. \tag{1.19}$$

Here a critical difference emerges between Bartik and using the shares as instruments: using the time-invariant industry shares as instruments necessitates using time-varying coefficients, as otherwise the predicted effect of a given industry on growth is fixed, and would be subsumed by the location fixed effects. Bartik puts the time-variation into the calculation

of the instrument. To see this, compare the first-stage using z_{2l}^0 as an instrument:

$$x_{lt} = \tilde{\tau}_l + \tilde{\tau}_t + z_{2l}^0 \tilde{\gamma} + u_{lt}. \tag{1.20}$$

Formally, Bartik and industry shares as instruments are only equivalent if the coefficient on z_{l2}^0 ($\tilde{\gamma}$) can be transformed into γ: $\gamma = \tilde{\gamma}/\Delta_{g,t}$. This will only hold generally if $\tilde{\gamma}$ is time-varying. Intuitively, Bartik allows a high service share to predict high growth in one period, but low growth in a different period, whereas the simplest form of using industry shares as instruments forces a high service share to always predict high (or low) growth.

To recover the equivalence between Bartik and using shares as instruments in the panel setting, we can interact industry shares with time fixed effects. Anticipating a point that we return to later, this approach means having many more instruments than in Bartik. Bartik is a single instrument whereas interacting T time periods and K industries gives $T \times K$ instruments.

Since the industry shares are time-invariant, the approach is similar to a differences-in-differences estimator. Here, the size of the policy is measured by the dispersion in national industry growth, $\Delta_{g,t} = \bar{g}_{2t} - \bar{g}_{1t}$, and the exposure to the policy is given by z_{2l}^0. And, because the industry shares are time invariant, $\tilde{\gamma}_t$ can only be estimated relative to a base period. This gives the familiar estimating equation:

$$y_{lt} = \alpha_l + \alpha_t + x_{lt}\beta + \epsilon_{lt} \tag{1.21}$$

$$x_{lt} = \tau_l + \underbrace{(\tau_t + \bar{g}_{1t}\gamma)}_{\tilde{\tau}_t} + z_{l2}^0 \underbrace{\Delta_{g,t}\gamma}_{\tilde{\gamma}_t} + u_{it} \tag{1.22}$$

$$= \tau_l + \tilde{\tau}_t + z_{l2}^0 \underbrace{\Delta_{g,t}\gamma}_{\tilde{\gamma}_t} + u_{lt} \tag{1.23}$$

$$= \tau_l + \tilde{\tau}_t + \sum_{s \neq 0} z_{l2}^0 \tilde{\gamma}_s 1(t = s) + u_{lt}. \tag{1.24}$$

This setup is familiar to the now-standard difference-in-difference setup with continuous treatment exposure, and non-parametrically estimated exposure. Cross-sectionally, some groups are exposed more or less to a treatment, which is measured by z_{2l}^0. to test for parallel trends.

REMARK 1.1. *In a panel setting, to maintain the analogy to difference-in-difference, the regression should include place and time fixed effects. A key difference compared to the standard setting is that a typical DinD will have a well-defined "shock" that treats one group vs. the other in one time period going forward, while the Bartik instrument generates a continuously varying treatment that*

can change over the sample.[4] *For example, there is a shock in 1970-1980, 1980-1990, and etc., of varying size across groups. This continuous variation in the treatment makes testing for parallel trends less direct. We return to this issue in section 3.4.*

In many applications, researchers are faced with the choice about whether to allow z_{lk} to vary over time, or fix it in an initial period. Let the initial period shares be denoted by z_{lk}^0. To see the need for fixing industry shares, consider the expression for next period's industry shares in the two-industry case, denoted by z_{lk}^1 and let g_{lk}^0 be period 0 growth rates:

$$z_{l1}^1 = \frac{g_{l1}^0 z_{l1}^0}{g_{l1}^0 z_{l1}^0 + g_{l2}^0 z_{l2}^0}; \; z_{l2}^1 = \frac{g_{l2}^0 z_{l2}^0}{g_{l1}^0 z_{l1}^0 + g_{l2}^0 z_{l2}^0}. \tag{1.25}$$

Note that the Bartik instrument is used due to concerns that either $\text{Cov}(g_{l1}^0, \epsilon_l^0) \neq 0$ or $\text{Cov}(g_{l2}^0, \epsilon_l^0) \neq 0$. (Or, more specifically, that $\tilde{g}_{l1}^0 = g_{l1}^0 - g_1^0$ is correlated with ϵ_l^0). If the ϵ_i^t are not independent over time (i.e., $\text{Cov}(\epsilon_l^0, \epsilon_l^1) \neq 0$), then updating the weights can induce a correlation between the instrument and the error term, rendering the instrument invalid.

REMARK 1.2 (General Industry Shares based on Initial Condition). *Generally, it is worth noting that this implies recursively:*

$$z_{lk}^t = z_{lk}^0 \prod_{s=0}^{t-1} \pi_{lk}^s \tag{1.26}$$

where $\pi_{lk}^s = \frac{g_{lk}^s}{x_l^s}$ is the relative growth in industry k in location l in period s. Hence, we can write the first-stage estimand as

$$\mathbb{E}(X|\mathbf{Z}^0) = \mathbf{Z}^{0\prime}\mathbb{E}(\prod_{s=0}^{t-1} \pi^s \mathbf{G}_t | \mathbf{Z}^0) \tag{1.27}$$

where π^s is a $K \times K$ diagonal matrix with the π_k^s (relative growth of industry k in period s) along the diagonals.

As a result of this potential bias, we use the initial industry share for each location, and hold it fixed across time periods. Hence, the Bartik estimator will be calculated as

$$B_{lt} = \bar{g}_{1t} z_{l1}^0 + \bar{g}_{2t} z_{l2}^0.$$

[4]For example, a simple case would be a change in trade tariffs in period $t = 0$: manufacturing is more affected than services. Hence, for $t > 0$, $\tilde{\gamma}_s$ would be positive, and would be a valid instrument so long as industry shares affected x_{lt} only through x_{lt}.

1.3 Many industries, many time periods

We now present the case with multiple industries and time periods. As in the two industry case, our goal is to show the relationship between Bartik and using industry shares as instruments. Let there be K industries, T time periods and N locations, with k,t, and i denoting a particular industry, time or location. Extending the two industry case, we now have that employment growth in location i at time t is given by:

$$X_{it} = \sum_{k=1}^{K} G_{ikt} Z_{ikt} \tag{1.28}$$

where G_{ikt} is the location-industry-time growth rate, and Z_{ikt} are the industry shares such that $\sum_{k \in \mathcal{I}} Z_{ikt} = 1, \forall i, t$.

We first derive an expression of the estimator using the Bartik instrument. Define

$$\hat{B}_{it} = \sum_{k \in \Sigma} \overline{G}_{kt} Z_{ikt}, \tag{1.29}$$

where $\overline{G}_{kt} = N^{-1} \sum_i G_{ikt}$ is the average industry growth rate for industry k in time period t.[5] The two stage least squares system of equations is:

$$Y_{it} = W_{it}\alpha + X_{it}\beta + \epsilon_{it} \tag{1.30}$$

$$X_{it} = W_{it}\tau + \hat{B}_{it}\gamma + u_{it}, \tag{1.31}$$

where W_{it} is a $1 \times L$ vector of controls. Typically in a panel context, W_{it} will include location and year fixed effects, while in the cross-sectional regression, this will simply include a constant. It may also include a variety of other variables. Let $n = N \times T$, the number of location-years. For simplicity, let $\mathbf{Y}_n$ denote the $n \times 1$ stacked vector of Y_{it}, $\mathbf{W}_n$ denote the $n \times L$ stacked vector of W_{it} controls and $\mathbf{X}_n$ denote the $n \times 1$ stacked vector of X_{it} and $\hat{\mathbf{B}}_n$ denote the stacked vector of $\hat{B}_{it}$. Denote $\mathbf{P_W} = \mathbf{W_n}(\mathbf{W_n}'\mathbf{W_n})^{-1}\mathbf{W_n}'$ as the $n \times n$ projection matrix of $\mathbf{W_n}$, and $\mathbf{M_W} = \mathbf{I}_n - \mathbf{P_W}$ as the annhilator matrix. Then, because this is an exactly identified instrumental variable our estimator is

$$\hat{\beta}_{\text{bartik}} = \frac{(\mathbf{M_W B}_n)'\mathbf{Y_n}}{(\mathbf{M_W B}_n)'\mathbf{X_n}}. \tag{1.32}$$

We now consider the alternative approach of using industry shares as instruments. The

[5] We return to leave-one-out below.

two-equation system is:

$$Y_{it} = W_{it}\alpha + X_{it}\beta + \epsilon_{it} \tag{1.33}$$

$$X_{it} = W_{it}\tau + Z_{it}\gamma_t + u_{it}, \tag{1.34}$$

where Z_{it} is a $1 \times K$ row vector of industry shares, and γ_t is a $K \times 1$ vector, and, reflecting the lessons of previous section, the t subscript allows the effect of a given industry share to be time-varying. In matrix notation, we write

$$\mathbf{Y_n} = \mathbf{W_n}\alpha + \mathbf{X_n}\beta + \epsilon_n \tag{1.35}$$

$$\mathbf{X_n} = \mathbf{W_n}\tau + \tilde{\mathbf{Z}}_n\gamma_T + u_n, \tag{1.36}$$

where γ_T is a stacked $1 \times (T \times K)$ row vector such that

$$\gamma_T = \left[\gamma_1 \cdots \gamma_T\right], \tag{1.37}$$

and $\tilde{\mathbf{Z}}_n$ is a stacked $n \times (T \times K)$ matrix such that

$$\tilde{\mathbf{Z}}_n = \left[\begin{array}{ccc} \mathbf{Z}_n \odot 1_{t=1} & \cdots & \mathbf{Z}_n \odot 1_{t=T} \end{array}\right], \tag{1.38}$$

where $1_{t=T}$ is an $n \times K$ indicator matrix equal to one if the nth observation is in period t, and zero otherwise. $\odot$ indicates the Hadamard product, or pointwise product of the two matrices. Let $\mathbf{Z}_n^{\perp} = \mathbf{M_W}\tilde{\mathbf{Z}}_n$ and $\mathbf{P}_{Z^{\perp}} = \tilde{\mathbf{Z}}_n^{\perp}(\tilde{\mathbf{Z}}_n^{\perp\prime}\tilde{\mathbf{Z}}_n^{\perp})^{-1}\tilde{\mathbf{Z}}_n^{\perp\prime}$. Then, the 2SLS estimator is

$$\hat{\beta}_{\text{2SLS}} = \frac{\mathbf{X_n}'\mathbf{P}_{Z^{\perp}}\mathbf{Y_n}}{\mathbf{X_n}'\mathbf{P}_{Z^{\perp}}\mathbf{X_n}}. \tag{1.39}$$

Alternatively, using the $\mathbf{Z}_n$ as instruments, the GMM estimator is:

$$\hat{\beta}_{\text{GMM}} = \frac{\mathbf{X_n}'\mathbf{M_W}\tilde{\mathbf{Z}}_n\Omega^{-1}\tilde{\mathbf{Z}}_n'\mathbf{M_W}\mathbf{Y_n}}{\mathbf{X_n}'\mathbf{M_W}\tilde{\mathbf{Z}}_n\Omega^{-1}\tilde{\mathbf{Z}}_n'\mathbf{M_W}\mathbf{X_n}}, \tag{1.40}$$

where Ω is a $(K \times T) \times (K \times T)$ weight matrix.

We now turn to the relationship between Bartik and using industry shares as instruments. To show the relationship between GMM and Bartik, it is helpful to write $\mathbf{B}_n$ in terms of the underlying industry shares and growth rates. Define a row vector $\tilde{\overline{\mathbf{G}}}$ which has an analogous structure to γ_T. We do this in two steps. First, let $\overline{G}_t$ be a $1 \times K$ row vector where the k^{th} entry is $\overline{G}_{kt}$. Second, stack the $\overline{G}_t$ to have the following row vector

$$\tilde{\overline{\mathbf{G}}} = \left[\overline{G}_1 \cdots \overline{G}_T\right], \tag{1.41}$$

where this vector is $1 \times (K \times T)$. Then

$$\mathbf{B}_n = \tilde{\mathbf{Z}}_n \tilde{\bar{\mathbf{G}}}'. \tag{1.42}$$

PROPOSITION 1.1. *If* $\Omega^{-1} = \tilde{\bar{\mathbf{G}}}' \tilde{\bar{\mathbf{G}}}$, *then* $\hat{\beta}_{GMM} = \hat{\beta}_{bartik}$.

Proof. See Appendix A. □

This proposition shows that Bartik is equivalent to doing GMM with industry shares as instruments when the inverse of the weight matrix is given by $\tilde{\bar{\mathbf{G}}}' \tilde{\bar{\mathbf{G}}}$.

Our framework also shows why—-starting with Autor and Duggan (2003)—the literature has often used a leave-one-out estimator of the industry growth rates to construct the Bartik instrument. In our notation,

$$\mathbf{B}_n^{lo} = \frac{N}{N-1} \tilde{\mathbf{Z}}_n \tilde{\bar{\mathbf{G}}}' - \frac{1}{N-1} \mathbf{X_n}, \tag{1.43}$$

where this expression is the same as $\mathbf{B}_n$ except that the own-location growth rate is subtracted off. Specifically, it is possible to derive an expression for the finite sample bias that emerges from including own-location in the computation of the growth rates.[6]

PROPOSITION 1.2. *In finite samples, the difference between the Bartik estimator of* β *and the true* β *is given by:*

$$\hat{\beta}_{bartik} - \beta = \frac{\tilde{\bar{\mathbf{G}}} \tilde{\mathbf{Z}}_n' \mathbf{M_W} \epsilon_n}{\tilde{\bar{\mathbf{G}}} \tilde{\mathbf{Z}}_n' \mathbf{M_W} \mathbf{X}_n} \tag{1.44}$$

where

$$\tilde{\bar{\mathbf{G}}} \tilde{\mathbf{Z}}_n' \mathbf{M_W} \epsilon_n = N^{-1} \sum_i Z_i G_i' M_{ii} \epsilon_i. \tag{1.45}$$

Proof. See Appendix A. □

Since G_i and ϵ_i are correlated, there will be bias in estimation when N is not sufficiently large. The leave-one-out mean estimator solves this problem directly, and is analogous to the jackknife estimator in JIVE (Angrist, Imbens, and Krueger (1999)).

[6] Note that in the presence of leave-one-out, the equivalence between Bartik and GMM no longer holds.

2 Estimation

So far we have established that using the Bartik *instrument* is equivalent to using industry shares as an instrument. Hence, what is distinctive about the Bartik instrument is not as an *instrument* per se, but as an *estimation approach* to dealing with a high-dimensional first stage when the endogenous variable has an inner product structure. Put differently, to call it the Bartik instrument is a slight misnomer, and it in fact could be called the Bartik approach.

We first explain the sense in which using industry shares as instruments generates a high-dimensional first stage. In the traditional two-stage estimator set-up, the generic first-stage estimation problem is to estimate $\mathbb{E}_l[x|Z]$. If a researcher estimates this conditional expectation using least squares with a vector of industry shares in the first-stage, the estimation would be very noisy if Z is high-dimensional and might lead to bias in the two-stage estimator. For example, a state-level analysis might have 50 (or 51) states and there are 312 4-digit NAICS industries. That is, in a single cross-section there would be more instruments than observations. In this case, even once the number of instruments is reduced to match the number of states, the high degree of over-identification instruments would generate substantial bias. As we emphasized above, moving to a panel setting does nothing to alleviate this problem because the industry shares are interacted with period dummies and so the number of instruments scales with the number of periods.

We now explain how Bartik uses the inner product structure of the endogenous variable to do dimension reduction. Given the inner product structure of x in the Bartik setting, the first-stage estimation problem can be rewritten as follows:

$$\mathbb{E}_l[x|Z] = \mathbb{E}_l[Z'G|Z] = Z'\mathbb{E}_l[G|Z].$$

In this case, instead of estimating $\mathbb{E}_l[x|Z]$, a researcher would need to estimate $\mathbb{E}_l[G|Z]$ and then pre-multiply it by Z. This estimation problem is still very high-dimensional. The Bartik instrument avoids the high-dimensional first-stage by making the following approximation:

$$\mathbb{E}_l[G|Z] \approx \mathbb{E}_l[G].$$

Then, and to return to the 4-digit industry example, rather than having 312 instruments, the researcher constructs a single instrument:

$$B = Z'\hat{\mathbb{E}}[G].$$

In this case, $\hat{\mathbb{E}}_l[G]$ is a valid estimator for $\mathbb{E}_l[G|Z]$, but just not necessarily the most efficient.

Substantively, this Bartik approximation means that the expectation of the growth rate of a particular industry does not depend on a location's industry composition. It would be

consistent with the identifying assumption of Bartik to allow for inter-industry spillovers through input-output linkages, i.e., the restaurant industry might have a different growth rate in locations with a large and small entertainment industry presence because restaurant industry demand comes primarily from the entertainment industry. Similarly, it is consistent with the Bartik assumption to allow for the effects of a shock to depend on levels, i.e., it might be that there is curvature in the location-industry production function and a given shock has a larger (or smaller) effect in locations where the industry is more prominent.

There are potentially many ways to efficiently estimate $\mathbb{E}_l[G|Z]$. Surprisingly, however, the Bartik estimator does extremely well in our application compared to other potential alternatives, in large part due to the fact that industry means explain almost 50% of the overall location-industry variance (see below for more details). Hence, simple means are a very reasonable approximation. In unreported results, we attempted other approximations to $\mathbb{E}_l[G|Z]$ but none worked any better than Bartik.

Since this process constructs a generated instrument to be used in a just-identified two-stage estimation procedure, there is no impact on the asymptotic distribution of the estimator. See Section 6.1.2 in Wooldridge (2002, pg. 117). This implies that under standard asymptotics, any improvement in estimating $\mathbb{E}[x|Z]$ by estimating $\mathbb{E}_l[G|Z]$ more efficiently will only improve our results in finite samples.

2.1 Industry and location bins: theoretical considerations

An important issue that arises in practice is how finely to divide industry and location bins. So far we have assumed a sampling frame in which it is not possible to discuss this question. The reason is that we have assumed that each location contains information on all industries. Hence, implicitly, as we divide industries and locations more finely, we get additional data on the additional industries and locations (and they are all independent), so there is no reason to not divide them arbitrarily finely. Here we consider an alternative sampling frame where it is possible to discuss this issue.

To be able to discuss the choice of how finely to divide locations (the set-up for industries would be analogous), we consider a sampling frame where the number of locations with a given industry is fixed as the sample size grows. We can motivate this frame by imagining that there are a fixed number of firms in a particular industry, so as we divide locations more finely we get no more information about the industry growth rates. Formally, let there be N locations, T time periods, and K industries. For each industry k, there are N_k locations where growth rates are observed. A simple way to envision this is that there are a number of individuals who are employed, and they are dispersed across locations in finite quantity. Hence, as N grows with K fixed, there are more and more cities with individuals in industry k, and so $N_k \to \infty$.

Another way of thinking about the fineness of industry and location bins is to think about the relationship between the assumptions at different levels of aggregation. Here we show two straightforward results. First, if Bartik is valid at a finer level of aggregation, then it is also valid at a coarser level. Second, the converse is not necessarily true, but we state assumptions that make this hold.

DEFINITION 2.1. *Let Z_i^1 and Z_i^2 be $K_1 \times 1$ and $K_2 \times 1$ vectors defining industry shares at different levels of aggregation ($K_1 < K_2$). We assume that there exists a surjective function f such that for each $k \in [1, \ldots, K_2]$, $f(k) \in [1, \ldots, K_1]$. Moreover, $\sum_{ks.t.f(k)=j} Z_k^2 = Z_j^1$.*

PROPOSITION 2.1. *If $\mathbb{E}(Z_i^2 \epsilon_i) = 0$, then $\mathbb{E}(Z_i^1 \epsilon_i) = 0$.*

Proof. Note that $\mathbb{E}(Z_{ik}^1 \epsilon) = \mathbb{E}(\sum_{j\text{s.t.}f(j)=k} Z_{ij}^2 \epsilon_i) = \sum_{j\text{s.t.}f(j)=k} \mathbb{E}(Z_{ij}^2 \epsilon_i) = 0$. □

REMARK 2.1. *If $\mathbb{E}(Z_i^1 \epsilon_i) = 0$, this does not necessarily imply that $\mathbb{E}(Z_i^2 \epsilon_i) = 0$. Note that $\sum_{js.t.f(j)=k} \mathbb{E}(Z_{ij}^2 \epsilon_i) = 0$ does not necessarily imply that $E(Z_{ij}^2 \epsilon_i) = 0$. However, it would require a particular knife edge case for this to hold that is unlikely under most empirical settings, as this combination of covariances would need to equal zero across all $k \in [1, \ldots, K_1]$.*

The theoretical results thus suggest that there is no strong reason to not arbitrarily divide industries as finely as possible.

2.2 First stage coefficient

A natural intuition is that the first stage coefficient in Bartik should be one. Here we show that the first-stage coefficient will be one (in the limit) when the Bartik approximation is exact.

Note that we can write $G_i = \bar{G} + u_i$, where u_i is a $K \times 1$ vector. Similarly, $B_i = Z_i' \bar{G}$.

$$\begin{aligned} Var(X_i) = Var(Z_i' G_i) &= Var(Z_i' \bar{G} + Z_i' u_i) && (2.1) \\ &= Var(Z_i' \bar{G} + 2Cov(Z_i' \bar{G}, Z_i' u_i) + Var(Z_i' u_i), && (2.2) \end{aligned}$$

In the limit as the number of industries grows large, the first stage coefficient is:

$$\begin{aligned} \hat{\gamma} &= \frac{Cov(B_i, X_i)}{Var(B_i)} && (2.3) \\ &= 1 + \frac{Cov(Z_i' \bar{G}, Z_i' u_i)}{Var(Z_i' \bar{G})}. && (2.4) \end{aligned}$$

Hence, whether the first stage coefficient is 1 depends on the properties of $Cov(Z_i' \bar{G}, Z_i' u_i)$. One sufficient condition for $Cov(Z_i' \bar{G}, Z_i' u_i) = 0$ is that $\mathbb{E}[G|Z] = \mathbb{E}[G]$.

PROPOSITION 2.2. *Suppose that G and Z are independent which implies that $\mathbb{E}[G|Z] = \mathbb{E}[G]$. Then $\mathbb{E}[\hat{\gamma}] = 1$.*

Proof.

$$
\begin{aligned}
Cov(Z_i'\bar{G}, Z_i'u_i) &= \mathbb{E}[Z_i'\bar{G}Z_i'u_i] - \mathbb{E}[Z_i'\bar{G}]\mathbb{E}[Z_i'u_i] && (2.5)\\
&= \mathbb{E}[Z_i'\bar{G}Z_i'(G_i - \bar{G})] - \mathbb{E}[Z_i'\bar{G}]\mathbb{E}[Z_i'(G_i - \bar{G})] && (2.6)\\
&= \mathbb{E}[Z_i'\bar{G}(G_i - \bar{G})'Z_i] - \mathbb{E}[Z_i'\bar{G}]\mathbb{E}[Z_i'(G_i - \bar{G})] && (2.7)\\
&= \mathbb{E}[Z_i'\mathbb{E}[\bar{G}(G_i - \bar{G})'|Z_i]Z_i] - \mathbb{E}[Z_i'E[\bar{G}|Z_i]]\mathbb{E}[Z_i'E[(G_i - \bar{G})]|Z_i] && (2.8)\\
&= \mathbb{E}[Z_i'\mathbb{E}[\bar{G}(G_i - \bar{G})']Z_i] - \mathbb{E}[Z_i']E[\bar{G}]\mathbb{E}[Z_i']E[(G_i - \bar{G})]] && (2.9)\\
&= 0. && (2.10)
\end{aligned}
$$

The first line is the definition of covariance, the second line is the definition of u_i, the third line takes the transpose of a scalar, the fourth line is the law of iterated expectations, the fifth line is the assumptin that G and Z are independent, and the sixth follows from the fact that $\mathbb{E}[G_i - \bar{G}] = 0$ and $Cov(\bar{G}, G_i - \bar{G}) = 0$. □

In finite samples, this result does not necessarily hold. Moreover, the term $\frac{Cov(Z_i'\bar{G}, Z_i'u_i)}{Var(Z_i'\bar{G})}$ can be large if Bartik has low explanatory power.

We explore this further in Section 4 when we perform simulations.

3 Testing for confounds

So far we have emphasized that the key assumption of Bartik is the exogeneity of industry shares. We now present some simple descriptive results that illustrate how one might go about probing this assumption. The key challenges are that industry shares are high-dimensional, and that Bartik allows for a new shock in every period so that testing for pre-trends is nonstandard. While Bartik is used for many questions (and the generic approach is used in many setting besides industry-location), we focus on variables related to labor supply, since this is the canonical application.

3.1 Dataset

We use the 5% sample of IPUMS (Ruggles et al. (2015)) for 1980, 1990 and 2000 and we pool the 2009-2011 ACSs for 2010. We look at PUMAs and 3-digit IND1990 industries. In Appendix C we show all our results with states and 2 digit industries and results are very

similar.[7] To construct the remaining aspects of our dataset, we follow Autor and Duggan (2003). In the notation given above, our y variable is earnings growth, and X is employment growth. We use people aged 18 and older who report usually working at least 30 hours per week in the previous year. We fix industry shares at the 1980 values, and then construct the Bartik instrument using 1980 to 1990, 1990 to 2000 and 2000 to 2010 leave-one-out growth rates.

3.2 Why might Bartik work?

Our theoretical results emphasize that the Bartik instrument uses cross-industry variation in national growth rates. And, if there is no cross-industry variation, then the instrument has no power. A simple variance decomposition of the industry-location growth rates shows that indeed there is a national component of the industry growth rates.

Consider the following expression for the industry-location growth rates in location i and industry k in a particular time period:

$$g_{ik} = g_k + g_i + \epsilon_{ik}. \tag{3.1}$$

We can operationalize this equation as a regression where we include location and industry fixed effects. By computing the covariance of each estimated component with the overall growth rate, we can then ask how much of the variance of the industry-location growth rates reflect the common industry component, how much is a location component, and how much is in the interaction.[8]

Table 2 provides evidence that there is indeed a common industry component to growth rates. At the 3 digit level and using PUMAs we find that the industry component explains 15-20% of the variance of growth rates. Interestingly, the location component is quite small–explaining less than 5% of the variance of the industry location growth rates. Table A2 shows analogous statistics at the 2-digit level and using locations, and finds a similarly small role for location in explaining the industry-location growth rates, but a substantially larger for industry (the industry component is about 45% of the variance).

The table also emphasizes a point that we return to in our simulations: there are many zeros in the industry-location growth rates—at the 3 digit level and with PUMAs, about a quarter of the industry-location level observations are zeros. In contrast, Table A2 shows that at the state and 2 digit industry level less than 5% of the state-industry observations

[7] We have also looked at the other 2 possible combinations of location and industry, and results are again quite similar. There are 244 3-digit IND1990 industries and 91 such 2 digit industries. There are 543 PUMAs.

[8] Formally, : $\frac{Cov(g_{ik},\hat{g}_k)}{Var(g_{ik})}$ (the industry share); $\frac{Cov(g_{ik},\hat{g}_i)}{Var(g_{ik})}$ (the location share); and $\frac{Cov(g_{ik},\hat{\epsilon}_{ik})}{Var(g_{ik})}$ (the residual share). We have also considered versions where we compute the industry share as the national leave-one-out mean and compute the g_k component directly, and get similar answers.

are zeros.

3.3 Balance

While the previous section showed that there is an important national component to industry growth rates, our theoretical results emphasized that identification in Bartik comes from the exogeneity of industry shares. A natural way to begin to probe the plausibility of this assumption is to examine the relationship between industry composition and observable characteristics of a place.

Table 3 shows that observables in both levels and changes are closely related to the Bartik projection of industry composition. The left-hand side of the table shows the relationship between these three different values of Bartik and *levels* of observable characteristics of the PUMAs. The right-hand side of the table shows these relationships for *changes* in observable characteristics of places. Observable characteristics explain a large share of the variance in the Bartik instrument. In the first period, the R^2 is over 0.6. The characteristic that is most consistently related to the Bartik instrument is education. Thus, any trend that is correlated with these characteristics—for example, skill biased technological change, immigration, or rising female labor force participation—will be correlated with Bartik.

The relationship between the Bartik instrument and the observable characteristics is because industry composition is related to observable characteristics, and not because of properties of the growth rates. In Table 4, we show the results of an analogous exercise where we do dimension reduction on the 1980 industry shares using principal component analysis, which does not use any information in the subsequent growth rates. We then relate the first principal component of 1980 industry shares to time-varying observable characteristics of a location. We find, if anything, a tighter relationship between the first principal component of the 1980 industry shares and characteristics. Notably, the relationship between the levels and the characteristics does not fade over time.

There are two broad classes of reactions to this observation. First, if we think that we live in a selection on observables world, then we have measured the observables and so we can control for them. Second, we might think that we live in a selection on unobservables world, and then worry that the extent of selection on observables suggests how much we should worry about selection on unobservables.

Table 5 pursues the selection on observables logic and reports the results of the IV estimates of the inverse labor supply elasticity with and without controlling for observables. The main take-away is that the inverse elasticity estimates are sensitive to the inclusion of controls for observable differences of locations. The table shows the results of pooling data for three ten-year changes (to 1990, 2000 and 2010). The benchmark IV estimate of the inverse elasticity of labor supply is shown in column (6) and is 1.08. Columns (7) through (9)

report results when we control for observable characteristics. Controlling for observables attenuates estimates by over 20% (from 1.08 to 0.82). Even though the regression includes location fixed effects, it is when we control for *levels* of observable characteristics rather than *changes* that the attenuation occurs.

3.4 Pre-trends

Besides balance, another natural way to explore the validity of an instrument is to consider pre-trends. While looking at balance forced us to focus on observables, examining pre-trends allows us to say something the relationship between unobservables and Bartik. A notable feature of Bartik is that it allows for a new shock in every period so that it implies that we do not expect parallel trends to hold. Here we develop a simple procedure to test for parallel trends even in the face of a time-varying instrument and show that there is evidence of pre-trends using the Bartik instrument.

3.4.1 Why parallel trends might not hold, even if Bartik is a valid instrument

It is consistent with the validity of the Bartik instrument to find evidence of pre-trends. To see this, suppose we have two periods of data, $t = \{1, 2\}$ and the same set-up we have been using:

$$y_{t,i} = \alpha_0 + \beta X_{t,i} + \epsilon_{t,i} \tag{3.2}$$

$$X_{t,i} = \alpha_1 + \gamma B_{t,i} + \nu_{t,i}, \tag{3.3}$$

where $B_{t,i}$ is a valid instrument. Note that y is already in changes. I.e., y might be wage growth. Testing for pre-trends then amounts to asking whether $Corr(y_1, B_2) = 0$. That is, do places with a higher period 2 instrument have faster (or slower) wage growth in period 1?

To see why there might be pre-trends even if the Bartik instrument is valid, note that we can write:

$$Cov(y_1, B_2) = Cov(\alpha_0 + \beta\alpha_1 + \beta\gamma B_1 + \beta\nu_1 + \epsilon_{1,}, B_2). \tag{3.4}$$

Hence, $Cov(y_1, B_2)$ can be nonzero if $Cov(B_1, B_2)$ is nonzero. That is, if the Bartik instrument is correlated through time (because, for example, industry growth rates are correlated through time), then we will find evidence of pre-trends.

3.4.2 Adjusting for mechanical pre-trends

Mechanically, we ask whether the residuals from the second stage in the current period can be predicted by the values of Bartik in a future period. That is, we remove the part of wage growth that we would predict from the Bartik instrument. Formally, we compute

$$\tilde{y}_{1,i} = y_{1,i} - \beta\gamma B_1. \tag{3.5}$$

Then this adjustment purges the most mechanical reason for pre-trends, and we have:

$$Cov(\tilde{y}_1, B_2) = Cov(\alpha_0 + \beta\alpha_1 + \beta\nu_1 + \epsilon_{1,}, B_2). \tag{3.6}$$

3.4.3 Evidence from Bartik

Column (1) of Table 6 shows that there is evidence of pre-trends. Namely, we pool wage growth from 1980 to 1990 and 1990 to 2000, and regress it on Bartik constructed one period forward, so the Bartik constructed using 1990 to 2000, and 2000 to 2010. Column (1) shows that we can predict past wage growth using future values of the instrument.

There is reason to think that some of this relationship might be mechanical. Namely, the correlation between adjacent period values of the Bartik instrument is 0.4272. Hence, it might be that the future value of the instrument is simply correlated with the past good shocks that led to the wage growth.

Columns (2) through (5) of Table 6 show that after addressing the mechanical reason for correlation that values of the Bartik instrument are correlated over time, we still find evidence of pre-trends. The columns correspond to the residualization in columns (6)-(9) of table 5 and show that for various ways of residualizing earnings growth for that predicted by the Bartik instrument we can still predict past values of wage growth using future values of the Bartik instrument. Two aspects of the table are quantitatively notable. First, controlling for observable characteristics of the location does reduce the magnitude of the implied pre-trends. Second, however, the size of the coefficient is still large. For example, the reduced-form for the effect of Bartik on wage growth for the specification in column (5) is 0.2624.[9] Hence, the coefficient in column (5) of 0.0403 of future values of Bartik on past values of wage growth is large.

4 Simulations

We now show a number of simulations that illustrate a few of the points we have made. Appendix B provides details on the simulations and Table A1 shows the parameters. The

[9] This multiplies the main effect in in column (5) and column (9) of Table 5.

baseline parameters are chosen such that Bartik "works."

Table 7 summarizes the interesting simulations. The first point that emerges from considering the various simulations is the role of the relative number of industries and locations. In the baseline simulation, we have 300 locations and 10 industries, that is, the ratio of locations to industries is 30. In this case, IV appears to be median and mean unbiased. When we drop the number of locations to 20 (so that the ratio of locations to industries is 2), we find that IV appears to be median and mean biased. Figure 1 shows the continuous version of this point. A similar way of showing this point is to hold the number of locations constant and raise the number of industries. Figure 2 shows that as we increase the number of industries the mean estimate of θ drifts down, while the variance across simulation runs increases. Table 7 shows that when we reach 200 industries (so that the ratio of locations to industries is 1.5), that IV is median and mean biased.

In these simulations, it appears that when the ratio of locations to industries approaches 2 or 3 that there is bias, while in our empirical work above we considered 543 locations and 244 industries (and in the appendix we considered 51 states and 91 industries). While we have not yet designed a simulation that we feel fully captures the empirical setting, this suggests that the fine-ness of industry and location bins can play a large role.

The other interesting point that emerges from the simulations is that reducing the variance of common industry component has large effects. The table shows the effect of dropping the variance by a factor of 7 (from $\sigma_k^2 = 7$ to $\sigma_k^2 = 1$). These simulations leave classic traces of a weak instrument, in the sense that the IV estimates are incredibly unstable (the 2.5th to 97.5th percentile of $\hat{\theta}$ across the simulations is -7.9 to 9.0). Figure 3 provides the continuous version of these results.

We have also explored how varying other dimensions of our baseline simulation affect results, but there are no notable or interpretable results. (See figures A1 to A3).

To understand the identification result, we consider three alternative simulations. First, having the industry common component in the error term does not necessarily constitute a problem. We draw a set of random vectors R_l in a way analogous to the Z, interact these with G_k, and enter them in the error term. The row titled "G_k in ϵ (R)" shows that having the industry growth rates in the first stage error does not by itself constitute a problem. Second, having some function of the industry shares in the error does constitute a problem for identification. We can show this in a couple ways. By analogy to the previous simulation, we take the inner product of the local industry shares and the national industry components and enter them in the error term. The next row titled "G_k in ϵ (Z)" shows that this leads to enormous bias in IV. We also generate a random vector, ϵ_K, which is unrelated to the G_k, interact this vector with the industry shares and enter it in the error term. This

also leads to substantial bias.[10]

5 Instruments encompassed by our structure

We now discuss several other instruments that our encompassed by our framework. This list cannot be exhaustive, but illustrates the widespread applicability of our results.

5.1 Immigrant enclave instrument

The immigrant enclave instrument popularized by Altonji and Card (1991) is also a Bartik-like instrument. They are interested in the effects of immigration on native wages, and are concerned that the correlation between immigrant inflows and local economic conditions may confound their estimates. To instrument for immigrant inflows, they "use the fraction of immigrants in a city in 1970 to predict the change in the fraction of immigrants over the following decade" (pg. 219). More recently, Lewis (2011) uses this instrument and argues that "persistence of regional immigration patterns derives from new immigrants' preference to resettle with family—and much of U.S. immigration is "family-based"—or to be in a culturally familiar environment." (pg. 1048). He constructs the instrument by country of origin.

To fit our notation, let x_l denote the number of newly arriving immigrants in a location in a given interval. Let k denote one of K countries of origin and Let z_{lk} denote the share of people arriving from origin country k living in location l. Hence, $\sum_{l=1}^{L} z_{lk} = 1, \forall k$. In contrast, in the industry-location setting it is the sum over k that sums to one. Let g_k denote the number of people arriving from location k. Then $x_l = \sum_k z_{lk} g_k$. The instrument comes from lagging the z_{lk}. Once we lag z, say z^0_{lk} in some initial period, then let i_{lk} be the number of immigrants from origin country k arriving in destination l. Then define $g_{lk} = \frac{i_{lk}}{z^0_{lk}}$ to be the hypothetical flow of immigrants from k that would have to have occurred to have generated the extent of flows; this allows us to write $x_l = \sum_k z^0_{lk} g_{lk}$. Then rather than using the g_{lk} that makes this an identity, the researcher uses $g_k = \sum_l i_{lk} = \sum_l g_{lk} z^0_{lk}$. (This is analogous to in the industry-location setting weighting the g_{lk} by the z_{lk} to compute the g_k, rather than equal-weighting across locations). This emphasizes that researchers should be using leave-one-out to compute the g_k.

The type of endogeneity that this deals with is that locations that receive an unusually large volume of immigrants from a particular origin might be hit by a good shock. I.e., let $\tilde{g}_{lk} = g_{lk} - g_k$. Then the instrument addresses the possibility that $\tilde{g}_{lk}$ is correlated with the error term.

[10] Admittedly, to get this to work, we need to make the element-by-element variance quite large: 70.

As has been recognized, the identifying assumption is that the z are exogenous. That is, that the initial immigrant enclaves are not correlated with *trends* in labor demand. For example, it would be problematic if immigrants are more likely to be (successful) entrepreneurs and so immigrant enclaves tend to be faster-growing areas.

5.2 China shock

Autor, Dorn, and Hanson (2013) are interested in understanding the effect of the increase in imports from China on local labor market outcomes. They are concerned that "realized US imports from China...may be correlated with industry import demand shocks" (pg. 2128) and thus propose an instrument. They describe their empirical strategy as follows: they exploit "cross-market variation in import exposure stemming from initial differences in industry specialization and instrument for US imports using changes in Chinese imports by other high-income countries" (pg. 2121).

In our notation, the authors' endogeneous variable of interst, x_l, is the commuting zone-specific increase in worker exposure to imports from China. This can written in inner product form, where Z_l is commuting-zone specific industry composition (lagged by several years). Ideally, the authors would observe g_{lk}, or commuting-zone and industry specific growth in import competition from China. That is, ideally it would be possible to map the rise of the import of each good from China to the local labor market. Instead, they observe g_k, or the national growth in imports. Hence, $x_l = \sum_k z_{lk} g_k$.

The problem of not observing g_{lk} is that g_k might contain import growth that reflects supply shocks in location l. To get around the problem that they do not observe the g_{lk} (and thus cannot do a leave-one-out estimator of $\hat{g}_k$), they use growth in imports to *other* high income countries from China.

This strategy addresses the following endogeneity concern. Let g_k^o be the growth in imports from China experienced by high-income countries in industry k. Then define $\tilde{g}_{lk} = g_{lk} - g_k^o$ as the industry-location idiosyncratic component of the imports. Loosely speaking, we might worry that some locations are more affected by the rise of imports in industry k than other locations in ways that are not predicted by variation in z_{lk}, and that these differences are correlated with supply shocks. This empirical strategy addresses this concern.

The key exogeneity assumption is that the industry shares are exogenous with respect to shocks. For example, if there are other trends in labor demand (or supply) that are correlated with industry composition, then this would be a problem for the empirical strategy. By focusing on within manufacturing variation, Autor, Dorn, and Hanson (2013) address some concerns.

5.3 Bank lending relationships

Greenstone, Mas, and Nguyen (2015) are interested in the effects of changes in bank lending on economic activity during the Great Recession. They observe county-level outcomes, and loan origination by bank to each county. Hence, in our notation, let x_l be credit growth in a county, let z_{lk} be the share of loan origination in county l from bank k in some initial period, and let g_{lk} be the growth in loan origination in county l by bank k over some period. Then $x_l = \sum_k z_{lk} g_{lk}$.

The most straightforward Bartik estimator would compute $\hat{g}_{-l,k} = \frac{1}{L-1} \sum_{l' \neq l} g_{l'k}$. However, Greenstone, Mas, and Nguyen (2015) are concerned that there is spatial correlation in the economic shocks and so leave-one-out is not enough to remove mechanical correlations. One approach would be to instead leave out *regions*. Instead, they pursue a generalization of this approach and regress:

$$g_{lk} = g_l + g_k + \epsilon_{lk}, \tag{5.1}$$

where the g_l and g_k are indicator variables for location and bank. Then the $\hat{g}_l$ captures the change in bank lending that is common to a county, while $\hat{g}_k$ captures the change in bank lending that is common to a bank. To construct their instrument, they use $\hat{B}_l = \sum_k z_{lk} \hat{g}_k$, where the $\hat{g}_k$ comes from equation (5.1).

This approach addresses the endogeneity that comes through $\hat{g}_l + \hat{\epsilon}_{lk}$. I.e., it removes the part of loan origination growth that is common to location, and idiosyncratic to bank-location pairs.

The identifying assumption remains that the z_{lk} are exogenous. As Greenstone, Mas, and Nguyen (2015) recognize, one way the z_{lk} would be endogenous is if the pre-existing bank lending relationships are not randomly assigned. For example, it is possible that some banks only operate in locations that were particularly hard hit by the recession. One way this could happen is if a bank specialized in subprime mortgages and thus was concentrated in locations that were hit hard by the recession.

5.4 Simulated instruments

Currie and Gruber (1996a) are interested in the effect of Medicaid eligibility on infant mortality and low birth weight. They want to exploit the expansion of Medicaid eligibility, and look at aggregate outcomes. They emphasize that "the actual fraction eligible depends on economic and demographic characteristics of the state, which may also be correlated with birth outcomes" (pg. 1275). Hence, they "instrument the actual fraction eligible with a measure of the generosity of Medicaid in a state and year that depends only on the state's eligibility rules."

To see this instrument as a Bartik-like instrument, let x_l be the share of the population in state l eligible for Medicaid, and note that rather than k indexing *industries*, k indexes one of K discrete *types*, where the types are defined by eligibility criterion of the policy. Z_l is the Medicaid eligibility rules in state l for each type. Finally, $\mathbf{G}$ is the share of the population in each state that is of each type. In this case, for $\hat{\mathbb{E}}[G|Z]$, Currie and Gruber use the national shares of each type of person. (Note that their estimation is in a panel with state and time fixed effects, so in practice they exploit changes in Z).

There are two differences from the Bartik setting. First, future values of Z are not mechanically related to past values of G, and so they update the weights by allowing for changes in policy to generate variation in the instrument over time. Second, the Z's are not shares and do not sum to one. This means that even if the G were constant—i.e, $g_{lk} = \frac{1}{|K|}$—then the instrument could still have variation and be relevant.

The simulated instrument deals with correlation between the states economic and demographic characteristics and birth outcomes. Divide the shares into the national component and the local component: $g_{lk} = g_k + \tilde{g}_{lk}$, where $g_k = \frac{1}{L}\sum_{l=1}^{L} g_{lk}$ and $\tilde{g}_{lk} = g_{lk} - g_l$. This approach addresses the possibility that the error term is correlated with $\tilde{g}_{lk}$. That is, that states with unusually high shares of poor women are likely to have worse birth outcomes for reasons that are independent of Medicaid.

As Currie and Gruber (1996a) explain (pg. 1276), "these models treat state Medicaid policy as though it were exogenous to birth outcomes." That is, the identifying assumption is that the Zs are exogenous. For example, if Medicaid eligibility is a function of last period's share of low birth weight babies, then this would be problematic for the instrument. They partially address this by including state fixed effects and by controlling for some time-varying characteristics. As we discuss below, since this is a continuous difference-in-differences approach, a more "modern" approach to this would be to look at balance and pre-trends.

Identical to the Bartik setting, the justification for using a simulated instrument is that it is dealing with a very high-dimensional instrument set. As (Currie and Gruber (1996a, pg. 1275)) emphasize, the simulated instrument "can be thought of as a convenient parameterization of legislative differences....in different states and years." That is, just as industry shares constitute a high-dimensional set of instruments, the details of eligibility also constitute a high-dimensional set of instruments. So the simulated instrument reduces the dimensionality of this instrument. Similarly, they recognize that the weights are in some sense arbitrary (pg. 1275, note 9): "the use of a national sample does not affect the consistency of our estimates, since any set of weights will still yield a measure that is a function only of state rules. "

5.5 Market size and demography

5.6 The "judge" instrument

6 Summary: recommendations for practice

This note develops a formal econometric structure to study the Bartik instrument. Developing such a structure is valuable because the Bartik instrument is widely used but not fully understood. We summarize this paper by the implications for practice that our structure delivers:

- The argument about exogeneity is in terms of the industry shares, and not growth rates.
- Use the initial period shares – don't update.
- Make sure to use leave-one-out means.
- Check to see how "filled" industries are – don't go too fine-grained in industry cuts if there aren't many cities covering the means. Check this. You want $N_k \propto N$.
- Balance tests– check for confounders using initial composition.
- Try looking for pre-trends after partialling out the direct effects of previous values of Bartik.

We wish to emphasize that while some of these recommendations may seem obvious parts of the applied microeconomics toolkit, we are struck by how rarely, if at all, they are used in the context of Bartik instruments.

References

Acemoglu, Daron and Joshua Linn. 2004. "Market Size in Innovation: Theory and Evidence from the Pharmaceutical Industry." *Quarterly Journal of Economics* 119:1049–1090.

Altonji, Joseph G. and David Card. 1991. "The Effects of Immigration on the Labor market Outcomes of Less-skilled Natives." In *Immigration, Trade and the Labor Market*, edited by John M. Abowd and Richard B. Freeman. University of Chicago Press, 201–234.

Angrist, Joshua D., Guido W. Imbens, and Alan B. Krueger. 1999. "Jackknife Instrumental Variables Estimation." *Journal of Aplied Econometrics* 14:57–67.

Autor, David, David Dorn, and Gordon Hanson. 2013. "The China Syndrome: Local Labor Market Effects of Import Competition in the United States." *American Economic Review* 103 (6):2121–2168.

Autor, David and Mark Duggan. 2003. "The Rise in the Disability Rolls and the Decline in Unemployment." *Quarterly Journal of Economics* 118 (1):157–205.

Bartik, Timothy. 1991. *Who Benefits from State and Local Economic Development Policies?* W.E. Upjohn Institute.

Blanchard, Olivier and Lawrence Katz. 1992. "Regional Evolutions." *Brookings Papers on Economic Activity* 1992 (1):1–75.

Bound, John and Harry J. Holzer. 2000. "Demand Shifts, Population Adjustments, and Labor Market Outcomes during the 1980s." *Journal of Labor Economics* 18 (1):20–54.

Currie, Janet and Jonathan Gruber. 1996a. "Health Insurance Eligibility, Utilization, Medical Care and Child Health." *Quarterly Journal of Economics* 111 (2):431–466.

———. 1996b. "Saving Babies: The Efficacy and Cost of Recent Changes in the Medicaid Eligibility of Pregnant Women." *Journal of Political Economy* 104 (6):1263–1296.

Freeman, Richard B. 1980. "An Empirical Analysis of the Fixed Coefficient "Manpower Requirement" Mode, 1960-1970." *Journal of Human Resources* 15 (2):176–199.

Greenstone, Michael, Alexandre Mas, and Hoai-Luu Nguyen. 2015. "Do Credit Market Shocks affect the Real Economy? Quasi-Experimental Evidence from the Great Recession and 'Normal' Economic Times." Working paper.

Kovak, Brian. 2013. "Regional Efects of Trade Reform: What is the Correct Measure of Liberalization?" *American Economic Review* 103 (5):1960–1976.

Lewis, Ethan. 2011. "Immigration, Skill Mix, and Capital Skill Complementarity." *Quarterly Journal of Economics* 126:1029–1069.

Ruggles, Steven, Katie Genadek, Ronald Goeken, Josiah Grover, and Matthew Sobek. 2015. *Integrated Public Use Microdata Series: Version 6.0 [Machine-readable database]*. Minneapolis: University of Minnesota.

Wooldridge, Jeffrey M. 2002. *Econometric Analysis of Cross Section and Panel Data*. MIT Press.

Table 1: Literature

Aizer (2010, AER)	Domestic Violence
Allcott and Kenniston (2014, WP R& R at REStud)	
Altonji and Card (1991, NBER Chapter)	Immigration
Autor and Duggan (2003, QJE)	Public, Disability insurance
Autor, Dorn and Hanson (2013, AER)	Trade, local labor markets
Autor, Dorn, Hanson and Song (2014, QJE)	Trade, local labor markets
Baum-Snow and Ferrreira (2015, Handbook of Urban and Regional)	Urban
Beaudry, Green and Sand (2012, ECMA)	Macro- Labor
Bertrand, Kamenica and Pan (2015, QJE)	
Blanchard and Katz (1992, BPEA)	Macro - Labor
Bloom, Draca and Van Reenen (2016, REStud)	Trade and Productivity
Bound and Holzer (2000, JoLE)	Labor
Brunner, Ross and Washington (2011, Restat)	Political economy
Cadena and Kovak (2016, AEJ:Applied)	Labor
Card (2001, JoLE)	Immigration
Card (2009, AER)	Immigration
Chodorow-Reich and Wieland (2016, WP)	Macro-labor, Reallocation
Davis and Haltiwanger (2014, WP– Jackson Hole)	
Diamond (2016, AER)	Urban/Public
Dinerstein, Hoxby, Meer and Villanueva (2014, NBER Chapter)	Education
Gould, Weinberg and Mustard (2002, REStat)	
Greenstone, Mas and Nguyen (2015, WP–R&R at AEJ:Policy)	Finance, Macro
Guerrieri, Hartley and Hurst (2013, JPubE)	Urban, Public
Hagedorn, Karahan, Manovskii and Mitman (2016, WP)	Public, Macro-Labor
Juhn and Kim (1999, JoLE)	
Kovak (2013, AER)	Trade
Lewis (2011, QJE)	
Lin (2011, REStat)	
Luttmer (2005, QJE)	Public
Moretti (2013, AEJ: Applied)	
Nakamura and Steinsson (2014, AER)	Macro, Fiscal Multipliers
Nekarda and Ramey (2011, AEJ: Macro)	Macro,
Notowidigdo (2013, WP)	
Oberfield and Raval (2014, WP – R&R at ECMA)	Macro and IO
Ottaviano, Peri and Wright (2013, AER)	Trade
Saiz (2010, QJE)	Urban
Saks and Wozniak (2011, JoLE)	
Suarez Serrato and Zidar (2016, AER))	Public Economics

Table 2: Variance decomposition of industry-location growth rates (Puma 3-Digit)

	Industry	Location	Residual	Share of zeros
1990	.1896	.0356	.7748	.2675
2000	.2015	.0255	.773	.2561
2010	.1352	.0183	.8465	.2911

Table 3: (Puma) 1980 3-Digit Industry Share Bartik Instrument

	Levels			Changes		
	1990	2000	2010	1990	2000	2010
Male	-0.31***	-0.24***	0.02	0.04	0.10*	-0.12**
White	-0.08	0.04	-0.04	-0.07	-0.13***	0.01
Native Born	-0.32***	-0.14**	-0.27***	-0.13***	0.10*	-0.09*
12th Grade Only	0.07	0.52***	0.20	-0.36***	-0.59***	-0.22**
Some College	0.69***	0.93***	0.95***	-0.53***	-1.14***	-1.28***
Veteran	0.29***	-0.14	0.37**	0.15***	-0.19***	0.10**
# of Children	0.05	0.03	0.26***	-0.04	0.18***	-0.22***
Total Income	-0.63*	0.11	-0.80***	0.56***	0.47***	0.78***
Social Security Income	0.17	-0.04	0.35***	-0.03	-0.01	-0.07
5-Year Same State	-1.14***	-0.03	-0.29	0.71**	0.23	
R^2	0.66	0.42	0.27	0.62	0.30	0.43
F	104.45	42.08	20.80	85.78	17.32	35.31
p	0.00	0.00	0.00	0.00	0.00	0.00

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 4: (Puma) 1980 3-Digit Industry Share Principal Component 1

	Levels				Changes		
	1980	1990	2000	2010	1990	2000	2010
Male	-0.39***	-0.38***	-0.38***	0.02	0.24***	0.09*	
White	-0.18***	-0.11***	-0.10***	-0.22***	-0.22***	-0.08	
Native Born	-0.61***	-0.44***	-0.42***	-0.15**	-0.20***	-0.20***	
12th Grade Only	0.08*	0.29***	0.20***	-0.34***	-0.18	-0.90***	
Some College	0.26***	0.50***	0.46***	-0.11	-0.02	-1.05***	
Veteran	0.17*	-0.11	-0.09	-0.39***	-0.31***	-0.18***	
# of Children	-0.14***	-0.12***	-0.17***	-0.07	0.28***	0.06	
Total Income	1.91***	0.93***	0.54***	0.63***	0.26*	-0.13	
Social Security Income	-0.74***	-0.26***	-0.22***	0.01	0.10**	0.07*	
5-Year Same State	-0.21	-0.10	-0.05	0.36	0.76		
R^2	0.82	0.87	0.85	0.73	0.57	0.52	
F	213.22	259.03	229.78	121.45	66.23	59.52	
p	0.00	0.00	0.00	0.00	0.00	0.00	

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 5: Across-Time Puma Regressions with 3-Digit Industry bartik_1980

	OLS	First Stages				Second Stages			
	Δ Wage (1)	Δ Emp (2)	Δ Emp (3)	Δ Emp (4)	Δ Emp (5)	Δ Wage (6)	Δ Wage (7)	Δ Wage (8)	Δ Wage (9)
Δ Emp	0.46***					1.08	0.82***	1.17***	0.79***
Bartik (1980)		0.38***	0.39***	0.27***	0.33***				
Male			-0.07		0.20*		-0.24***		-0.30***
White			0.22		0.16		0.22*		0.13
Native Born			-0.04		-0.20		0.15		0.13
12th Grade Only			0.47***		0.14		-0.36***		-0.11
Some College			0.63***		0.66***		-0.71***		-0.58***
Veteran			0.04		-0.20		0.15*		0.20**
# of Children			0.76***		0.23*		-0.42***		-0.34***
Δ Male				0.05	0.08*			-0.01	-0.05*
Δ White				0.01	0.01			-0.07	-0.03
Δ Native Born				-0.05	-0.04			0.20***	0.12***
Δ 12th Grade Only				-0.19***	-0.23***			0.18***	0.18***
Δ Some College				-0.18*	-0.06			0.24***	0.08
Δ Veteran				-0.13**	-0.15**			0.21***	0.17***
Δ # of Children				-0.26***	-0.27***			0.11*	0.00
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Puma FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1,629	1,629	1,629	1,629	1,629	1,629	1,629	1,629	1,629
R-squared	0.90	0.73	0.77	0.79	0.80	0.78	0.88	0.78	0.90
F	.	.	.	.	.				
p-value	.	.	.	.	.		0.00	0.00	0.00

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 6: Regression of 1-Lagged Wage Growth Residuals on Bartik, Puma 3-Digit

		Residualized			
	Lag Wage Growth	No Controls	Levels	Changes	Levels+Changes
	(1)	(2)	(3)	(4)	(5)
bartik_1980	0.112**	0.0988***	0.0558***	0.0567***	0.0403***
	(0.0360)	(0.0125)	(0.0104)	(0.0106)	(0.0101)
N	1086	1086	1086	1086	1086

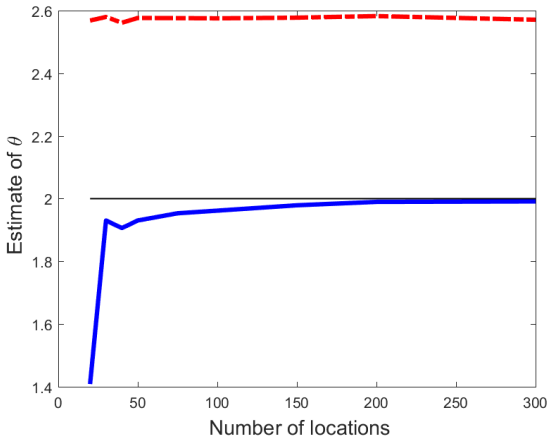
Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 7: Simulation results

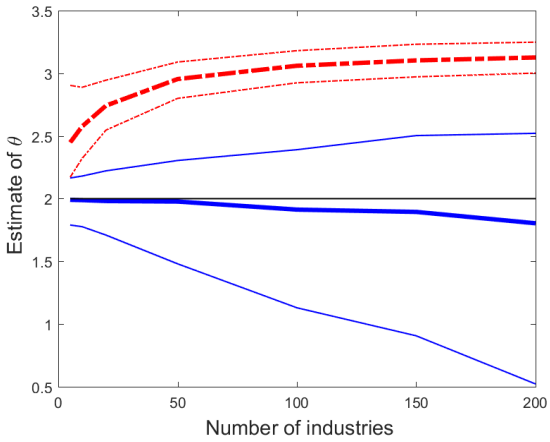
	OLS				IV			
	Mean($\hat{\theta}$)	Median ($\hat{\theta}$)	$\hat{\theta}_{2.5}$	$\hat{\theta}_{97.5}$	Mean($\hat{\theta}$)	Median ($\hat{\theta}$)	$\hat{\theta}_{2.5}$	$\hat{\theta}_{97.5}$
Benchmark	2.5760	2.5606	2.3321	2.8806	1.9927	2.0016	1.7686	2.1764
$L = 20$	2.5684	2.5472	2.0478	3.1621	1.4084	1.9365	-0.1614	2.5931
$K = 200$	3.1263	3.1262	3.0017	3.2485	1.8036	1.9053	0.5218	2.5212
$\sigma_k^2 = 1$	3.1030	3.1038	2.9735	3.2267	0.7719	1.8535	-7.8606	8.9783
G_k in ϵ (R)	2.5736	2.5638	2.2961	2.8884	1.9930	1.9992	1.7434	2.1963
G_k in ϵ (Z)	3.0628	3.0613	2.9821	3.1590	3.0008	3.0018	2.8482	3.1512
ϵ_k in ϵ (Z)	2.6803	2.6695	-0.8200	6.1420	2.2531	2.1527	-5.1831	10.6310

Figure 1: Bartik simulation: changing number of locations



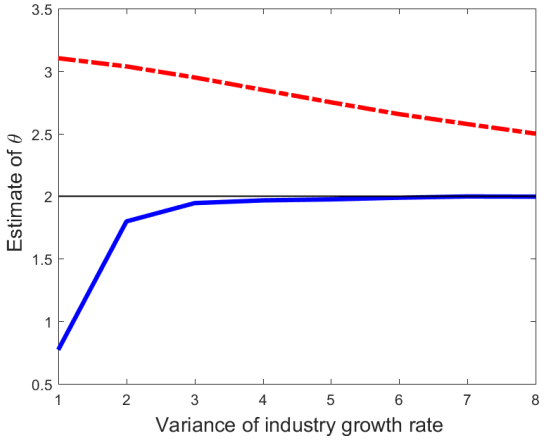
Notes: The blue line shows the TSLS estimates using Bartik. The red dashed line shows OLS. The thin black line shows the truth ($\theta = 2$). The thin lines show the 2.5 to 97.5th percentiles across the 1000 simulations at each point. The parameter values are as in Table A1, except that L varies. The benchmark number of locations is 300.

Figure 2: Bartik simulation: changing number of industries



Notes: The blue line shows the TSLS estimates using Bartik. The red dashed line shows OLS. The thin black line shows the truth ($\theta = 2$). The thin lines show the 2.5 to 97.5th percentiles across the 1000 simulations at each point. The parameter values are as in Table A1, except that K varies. The benchmark number of industries is 10.

Figure 3: Bartik simulation: changing variance of industry shocks



Notes: The blue solid line shows the TSLS estimates using Bartik. The red dashed line shows OLS. The thin black line shows the truth ($\theta = 2$). There are 1000 simulations at each point. The parameter values are as in Table A1, except that σ_k^2 varies. The benchmark value of σ_k^2 is 7.

A Omitted proofs

Proposition 1.1

Proof. Start with the Bartik estimator,

$$\hat{\beta}_{\text{bartik}} = \frac{(\mathbf{M_W B}_n)'\mathbf{Y_n}}{(\mathbf{M_W B}_n)'\mathbf{X_n}} \tag{A1}$$

$$= \frac{\mathbf{B}_n'\mathbf{M_W Y_n}}{\mathbf{B}_n'\mathbf{M_W X_n}} \tag{A2}$$

$$= \frac{\tilde{\bar{\mathbf{G}}}\tilde{\mathbf{Z}}_n'\mathbf{M_W Y}_n}{\tilde{\bar{\mathbf{G}}}\tilde{\mathbf{Z}}_n'\mathbf{M_W X}_n} \tag{A3}$$

$$= \frac{\mathbf{X}_n'\mathbf{M_W}\tilde{\mathbf{Z}}_n\tilde{\bar{\mathbf{G}}}'\tilde{\bar{\mathbf{G}}}\tilde{\mathbf{Z}}_n'\mathbf{M_W Y}_n}{\mathbf{X}_n'\mathbf{M_W}\tilde{\mathbf{Z}}_n\tilde{\bar{\mathbf{G}}}'\tilde{\bar{\mathbf{G}}}\tilde{\mathbf{Z}}_n'\mathbf{M_W X}_n}, \tag{A4}$$

where the second equality is algebra, the third equality follows from the definition of $\mathbf{B}_n$, and the fourth equality follows because $\mathbf{X}_n'\mathbf{M_W}\tilde{\mathbf{Z}}_n\tilde{\bar{\mathbf{G}}}'$ is a scalar. By inspection, if $\Omega^{-1} = \tilde{\bar{\mathbf{G}}}'\tilde{\bar{\mathbf{G}}}$, then $\hat{\beta}_{GMM} = \hat{\beta}_{bartik}$. □

Proposition 1.2

Proof. Note that if $\mathbf{B_n}$ is defined using a leave-one-out estimator for each location, this equivalence does not hold exactly. Instead, $\mathbf{B_n}$ can be written as

$$\mathbf{B}_n = \frac{N}{N-1}\tilde{\mathbf{Z}}_n\tilde{\bar{\mathbf{G}}}' - \frac{1}{N-1}\mathbf{X_n}. \tag{A5}$$

To see this, note that in this case,

$$B_{it}^{lo} = \sum_k \sum_i Z_{ikt}(N-1)^{-1}\sum_{j\neq i} G_{jkt} \tag{A6}$$

$$= \sum_k \sum_i Z_{ikt}(N-1)^{-1}(N\bar{G}_{kt} - G_{ikt}) \tag{A7}$$

$$= (N-1)^{-1}\left[N\sum_k\sum_i Z_{ikt}\bar{G}_{kt} - Z_{ikt}G_{ikt}\right] \tag{A8}$$

$$= (N-1)^{-1}N\sum_i\sum_k Z_{ikt}\bar{G}_{kt} - (N-1)^{-1}\sum_k Z_{ikt}G_{ikt} \tag{A9}$$

$$= (N-1)^{-1}N\sum_i\sum_k Z_{ikt}\bar{G}_{kt} - (N-1)^{-1}X_i. \tag{A10}$$

Hence, this implies that

$$\hat{\beta}_{\text{bartik}} = \frac{\left[\tilde{\bar{\mathbf{G}}}\tilde{\mathbf{Z}}_n' - N^{-1}\mathbf{X_n}'\right]\mathbf{M_W}\mathbf{Y}_n}{\left[\tilde{\bar{\mathbf{G}}}\tilde{\mathbf{Z}}_n' - N^{-1}\mathbf{X_n}'\right]\mathbf{M_W}\mathbf{X}_n}. \tag{A11}$$

Note that for sufficiently large N, $\hat{\beta}_{\text{bartik}} \approx \hat{\beta}_{\text{GMM}}$. This also highlights where the finite sample bias from failing to use the leave-one-out mean in $\overline{G}$ comes from. Note that

$$\hat{\beta}_{\text{bartik}} - \beta = \frac{\tilde{\bar{\mathbf{G}}}\tilde{\mathbf{Z}}_n'\mathbf{M_W}\epsilon_n}{\tilde{\bar{\mathbf{G}}}\tilde{\mathbf{Z}}_n'\mathbf{M_W}\mathbf{X}_n} \tag{A12}$$

and

$$\tilde{\bar{\mathbf{G}}}\tilde{\mathbf{Z}}_n'\mathbf{M_W}\epsilon_n = \sum_i \sum_j Z_i \overline{G}' M_{ij}\epsilon_j. \tag{A13}$$

For $i \neq j$, we have assumed independence, so bias would arise when $i = j$:

$$\begin{aligned}
\tilde{\bar{\mathbf{G}}}\tilde{\mathbf{Z}}_n'\mathbf{M_W}\epsilon_n &= \sum_i Z_i \overline{G}' M_{ii}\epsilon_i && \text{(A14)}\\
&= \sum_i Z_i N^{-1} \sum_j G_j' M_{ii}\epsilon_i && \text{(A15)}\\
&= N^{-1} \sum_i Z_i G_i' M_{ii}\epsilon_i. && \text{(A16)}
\end{aligned}$$

□

B Simulation details

B.1 Overview

We first set-up a simulation where Bartik works. Here is the notation:

- K: number of industries
- L: number of locations
- $g_k \sim N(0, \sigma_k^2)$: industry growth rates
- $g_{kl} \sim N(0, \sigma_{kl}^2)$: industry-location growth rates
- $g_l \sim N(0, \sigma_l^2)$: location growth rate
- $g_{lk}^{tot} = g_k + g_{kl} + g_l$: total growth observed in the industry-location
- G_l: the vector version of g_{lk}^{tot}
- To construct Z_l: draw independent standard normal variables, take the absolute values, and then normalize such that the Z's sum to one in each location
- $X_l = Z_l' G_l$: growth rate of "employment" in the location
- $y_l = 0.5 + \theta X_l + \epsilon_l$: per capita earnings growth in the location, where $\epsilon_l \sim N(0, \sigma_\epsilon^2)$ and ϵ_l is correlated with X_l through $Corr(g_l, \epsilon_l) = \rho$

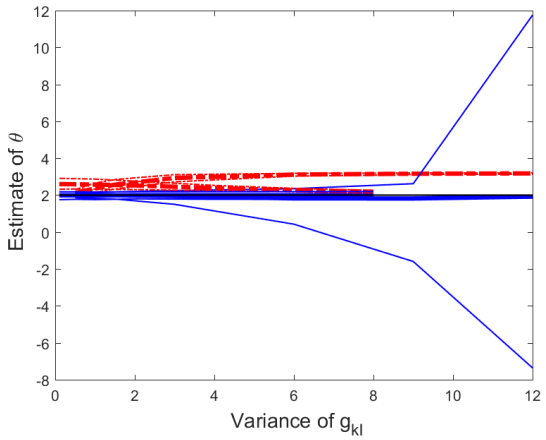
Table A1: Base simulation

Parameter	Value
K	10
L	300
σ_k^2	7
σ_{kl}^2	1
σ_l^2	1.5
θ	2
σ_ϵ^2	2
σ_z^2	1
ρ	0.6

We show simulations where we do the following:

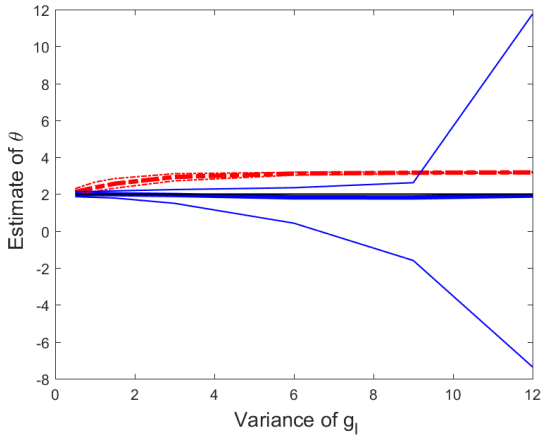
- Vary K, L, σ_k^2, σ_{kl}^2, σ_l^2 and σ_ϵ^2.
- Define: $y_l = 0.5 + \theta X_l + \epsilon_l + Z_l' G_k$, where G_k is the vectorized version of g_k

Figure A1: Bartik simulation: changing variance of industry-location shocks



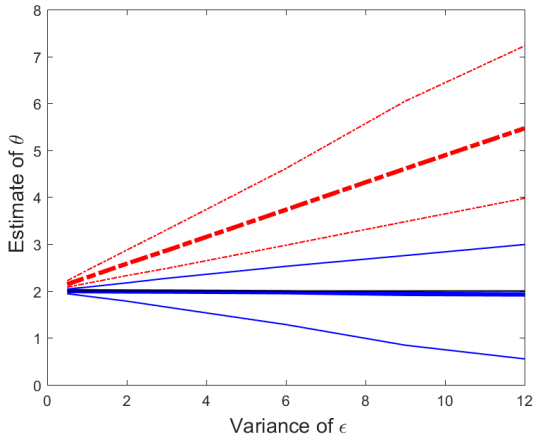
Notes: The blue solid line shows the TSLS estimates using Bartik. The red dashed line shows OLS. The thin black line shows the truth ($\theta = 2$). There are 1000 simulations at each point. The parameter values are as in Table A1, except that g_k varies. The benchmark value of σ^2_{kl} is 1.

Figure A2: Bartik simulation: changing variance of location shocks



Notes: The blue solid line shows the TSLS estimates using Bartik. The red dashed line shows OLS. The thin black line shows the truth ($\theta = 2$). The thin lines show the 2.5 to 97.5th percentiles across the 1000 simulations at each point. The parameter values are as in Table A1, except that σ^2_l varies. The benchmark value of σ^2_l is 1.5.

Figure A3: Bartik simulation: changing variance of error term



Notes: The blue solid line shows the TSLS estimates using Bartik. The red dashed line shows OLS. The thin black line shows the truth ($\theta = 2$). The thin lines show the 2.5 to 97.5th percentiles across the 1000 simulations at each point. The parameter values are as in Table A1, except that σ^2_ϵ varies. The benchmark value of σ^2_ϵ is 2.

C Additional tables

Table A2: Shares State 2-Digit

	Industry	Location	Residual	Share of zeros
1990	.4562	.0543	.4896	.0509
2000	.4756	.0662	.4582	.0269
2010	.4006	.0323	.5672	.0304

Table A3: (State) 1980 2-Digit Industry Share Bartik Instrument

	Levels			Changes		
	1990	2000	2010	1990	2000	2010
Male	-0.35*	-0.46**	-0.08	0.18	0.04	0.06
White	-0.14	0.06	-0.07	-0.03	0.08	0.24
Native Born	-0.46***	-0.34***	-0.25	-0.12	-0.12	-0.10
12th Grade Only	-0.02	0.23	0.73	-0.29*	-1.05***	0.36
Some College	0.18	0.43*	1.43**	-0.82***	-1.50***	-1.78**
Veteran	0.66*	0.40	-0.12	0.14	-0.12	0.22
# of Children	0.13	0.07	0.27	-0.07	0.11	-0.27
Total Income	0.43	0.20	-0.96*	0.04	0.20	0.39
Social Security Income	1.47*	0.35	0.52	0.12	-0.08	-0.23
5-Year Same State	-1.14	-0.14	-1.70***	-0.78	1.77	
R^2	0.74	0.63	0.47	0.63	0.67	0.65
F	22.83	9.15	8.51	29.30	12.55	20.11
p	0.00	0.00	0.00	0.00	0.00	0.00

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A4: (State) 1980 2-Digit Industry Share Principal Component 1

	Levels			Changes		
	1990	2000	2010	1990	2000	2010
Male	-0.27	-0.52	-0.00	-0.15	0.18	0.11
White	0.46***	0.51*	0.35*	-0.12	-0.06	-0.38*
Native Born	0.17	0.14	0.44**	0.59**	1.11***	0.65***
12th Grade Only	0.31*	0.14	-0.17	0.02	0.54	0.18
Some College	0.09	-0.28	-0.61	1.05*	0.53	1.87**
Veteran	-0.10	-0.07	-0.86	-0.56***	-0.63**	-0.72***
# of Children	-0.01	-0.03	-0.23	0.28	0.45*	0.48*
Total Income	-2.06	0.77	0.83*	0.31	1.43	0.39
Social Security Income	-2.19*	-0.29	0.04	-0.21	0.13	0.11
5-Year Same State	4.00***	1.91***	2.15***	-2.44**	-2.17	
R^2	0.85	0.75	0.79	0.77	0.58	0.63
F	59.34	16.38	19.15	23.51	7.52	8.63
p	0.00	0.00	0.00	0.00	0.00	0.00

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A5: Across-Time State Regressions with 2-Digit Industry bartik_1980

	OLS	First Stages				Second Stages			
	Δ Wage	Δ Emp	Δ Emp	Δ Emp	Δ Emp	Δ Wage	Δ Wage	Δ Wage	Δ Wage
Δ Emp	0.46***					0.95***	0.76***	2.52	1.18**
Bartik (1980)		0.70***	0.70**	0.19	0.36				
Male			-0.75		0.03		0.20		-0.25
White			2.56*		1.34		-0.87		-1.03
Native Born			-0.08		-0.58		0.32		0.46
12th Grade Only			0.65		0.29		-0.46*		-0.30
Some College			1.47*		1.10		-1.21***		-1.21*
Veteran			0.20		-0.07		0.36		0.33
# of Children			1.58***		0.30		-0.58*		-0.57
Δ Male				0.30	0.37			-0.82	-0.52*
Δ White				0.24*	0.23			-0.61	-0.24
Δ Native Born				0.10	-0.03			-0.28	-0.08
Δ 12th Grade Only				-0.43*	-0.32			1.04	0.43*
Δ Some College				-1.10**	-0.80			2.71	0.88
Δ Veteran				-0.26	-0.25			0.82	0.46*
Δ # of Children				-0.42***	-0.41**			0.73	0.15
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
State FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	153	153	153	153	153	153	153	153	153
R-squared	0.93	0.67	0.80	0.85	0.87	0.84	0.92	0.20	0.87
F	.	.	.	.	.				
p-value	.	.	.	.	.	0.00	0.00	0.00	0.00

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A6: Regression of 1-Lagged Wage Growth Residuals on Bartik, State -Digit

	OLS	Residualized			
	Lag Wage Growth	No Controls	Levels	Changes	Levels+Changes
bartik_1980	0.459***	0.140**	0.0306	0.0225	0.0133
	(0.129)	(0.0486)	(0.0331)	(0.0330)	(0.0315)
N	102	102	102	102	102

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$