WORKING PAPER · NO. 2024-12

Difference-in-Differences in the Marketplace

Robert Minton and Casey B. Mulligan

FEBRUARY 2024

Becker Friedman Institute
FOR ECONOMICS AT UCHICAGO

5757 S. University Ave.
Chicago, IL 60637
Main: 773.702.5599
bfi.uchicago.edu

DIFFERENCE-IN-DIFFERENCES IN THE MARKETPLACE

Robert Minton
Casey B. Mulligan

This paper grew out of a project of adding new chapters to Chicago Price Theory, in preparation for its second edition. Kevin Murphy's Price Theory lectures and further recommendations strongly influenced the preparation of this paper. We also appreciate comments from Alex Torgovitsky, Giuseppe Forte, Josh Gross, João Pugliese, and Alex Tordjman. The views in this paper are those of the authors and do not represent those of the Board of Governors of the Federal Reserve System or the Federal Reserve System or the National Bureau of Economic Research.

© 2024 by Robert Minton and Casey B. Mulligan. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Difference-in-Differences in the Marketplace
Robert Minton and Casey B. Mulligan
February 2024
JEL No. C21,D41,L11

ABSTRACT

Price theory says that the most important effects of policy and technological change are often found beyond their first point of contact. This appears opposed to econometric methods that rule out spillovers of one person's treatment on another's outcomes. This paper uses the industry model from price theory to represent the statistical concepts of treatments and controls. When treated and control observations are in the same market, the controls are indirectly affected by the treatment. Moreover, even the effect of the treatment on the treated reveals only part of the consequence for the treated of treating the entire market, which is often the parameter of interest. Marshall's Laws of Derived Demand provide a guide for empirical work: precise price-theoretic interpretations of the direct and spillover effects of a treatment, the quantitative relationships between them, and how they correspond to the scale and substitution effects emphasized in price theory.

Robert Minton
Federal Reserve Board
2001 C St NW
Office M-1703
Washington, DC 20037
robert.j.minton@frb.gov

Casey B. Mulligan
University of Chicago
Department of Economics
1126 East 59th Street
Chicago, IL 60637
and NBER
c-mulligan@uchicago.edu

I. Introduction

Price theory says that the most important effects of policy and technological change are often found beyond their first point of contact. This appears opposed to econometric methods that rule out "spillovers" of one person's "treatment" on another's outcomes. Our purpose here is not to discourage such methods, but rather to use price theory to help understand what they measure and how empirical findings can be applied to settings other than the ones where the measurement occurred.

This paper begins with a labor-market equilibrium example well-known in price theory's oral tradition, relating it to the difference-in-differences (DiD) method from econometrics. Section III then uses an extension of the industry model from price theory (Jaffe, Minton, Mulligan, & Murphy, 2019) to represent the statistical concepts of treatments and controls. When treated and control observations are in the same market, the controls are indirectly affected by the treatment. Moreover, even the effect of the treatment on the treated reveals only part of the consequence for the treated of treating the entire market, which is often the parameter of interest. We draw on Marshall's Laws of Derived Demand to provide precise price-theoretic interpretations of the direct and spillover effects of a treatment, the quantitative relationships between them, and how they correspond to the scale and substitution effects emphasized in price theory.

Our framework helps address a couple of misunderstandings about when spillover effects occur and how they affect the interpretation of DiD estimates. Interestingly, as the share treated falls, the direct effects of the treatment diminish more than the equilibrium effects do. Section IV shows how the substitution effect isolated by DiD can help construct estimates of the scale effect. Sections V and VI conclude with additional applications in which acknowledging equilibrium effects profoundly changes the interpretation of DiD estimates, including instances in which DiD estimates "have the wrong sign" due to market forces.

Econometric results on causal research designs, along with recent extensions in the literature, often rely on the assumption of "no spillovers."[1] "Spillovers" and "peer effects" are treated in microeconometrics as advanced, albeit interesting, topics that primarily arise when there are "externalities" (Angrist and Pischke 2008, Athey and Imbens 2017). Attempts to relax this assumption entail structure on how treatment spills over to the controls (Manski 1993)—a structure which could be economic or statistical. The statistical approach reviewed in Huber (2023) might allow spillovers from observations within predetermined clusters but not from observations outside those clusters (Sobel 2006, Hong and Raudenbush 2006, Hudgens and Halloran 2008). Our contribution aligns with the economic approach, which we view as lacking in the general frameworks more recently available in statistics. We provide closed-form results for interpreting quantity or price comparisons, showing how these estimates related to broader treatment effects on the entire market. Our approach focuses on spillovers mediated by market forces as opposed to

[1] See, for example, de Chaisemartin and d'Haultfoeuille (2020); Goldsmith-Pinkham, Sorkin and Swift (2020); and Borusyak, Hull and Jaravel (2022). Sometimes "no spillovers" is called "no interference" or is wrapped into the broader notion of the "stable unit treatment value assumption."

spillovers through externalities (such as the urban knowledge spillovers in Jacobs (1969) or spillovers of medical treatment in Miguel and Kremer (2004)).

The analysis of specific market-based spillovers is extensive and spans many fields. In urban economics, for example, Glaeser and Gottlieb (2009) assess the benefits of easy labor mobility across firms within cities. In labor economics, Monte, Redding, and Rossi-Hansberg (2018) find evidence that commuting is an important adjustment mechanism for localized labor demand shocks. Crépon et al. (2013) find that gains to unemployed job seekers of job placement assistance can be offset by displacement effects for those who did not receive the program. In development economics, Egger et al. (2022) find that transfer payments in one village can affect outcomes in nearby villages, although the market forces featured in our paper are not necessarily "general equilibrium" because they can occur in a single market. As discussed in our Section VII, public economics acknowledges that the introduction of state-specific cigarette taxes may affect the wholesale price of cigarettes faced by all states, a broader market response not captured by analyses comparing retail price changes of cigarettes in different states.

Our contribution to this research space, into which we have provided only a small glimpse, is a general equilibrium framework that researchers can apply to assess what market spillovers are present and how important they might be. While our approach can be used as a substitute for purely statistical models accounting for spillovers, it also has complementary elements. For instance, our results can provide useful insight for dividing observations into clusters within which spillovers are permitted from clusters where they are not, as discussed in Section IV.

II. A labor market illustration of equilibrium spillovers

From an input perspective, barbers today cut hair almost exactly as they did in the first half of the twentieth century: a chair, mirror, scissors, and sink. By all accounts, fully-scheduled barbers have hardly changed the number of haircuts they perform per hour. Meanwhile, other occupations have experienced dramatic productivity growth over the same time frame. For example, the number of bushels of corn produced per farmer has increased by an order of magnitude (U.S. Department of Agriculture, National Institute of Food and Agriculture 2014).

Given that the trends for inflation-adjusted wages of farmers and barbers are nearly the same despite their disparate productivity experiences, can we conclude that the causal effect of productivity on wages is essentially nil? That would appear to be the answer coming from the "difference-in-differences" statistical method.

Specifically, the DiD approach forms a "treatment group" as a sample of observations that were "exposed" to a "treatment," to be compared with a "control group" that was not "exposed." The treatment in our example is productivity growth. The occupation of farmer might be considered a treatment group because farmers became significantly more productive on their jobs. Barbers could serve as a control group because "haircutting has exhibited virtually no productivity improvement over a century." (Krueger 1991) The difference in the log of barbers' real wages now from a century ago is about 2, as is the difference for farmers. In its simplest form, the DiD method calculates the difference between two differences: one treatment difference and another

control difference. Here the DiD is essentially zero because the two occupations have similar real-wage growth. In other words, the DiD seems to show that even massive productivity growth of the amount experienced by farmers has a trivial effect on real wages. Conversely, if productivity is an important determinant of real wages, the DiD estimate would seem to present us with a puzzle.

The price theory solution to the puzzle is that occupation is a matter of choice. If barbers are to voluntarily cut hair, their real wage must somehow keep up with real wages of alternative occupations. That happens with a rising price of haircuts relative to corn. Through labor markets, the wage growth of barbers is largely determined by the productivity growth of other occupations. To put it another way, the DiD "correctly" shows that occupation-specific productivity growth has little occupation-specific effect on real wages, but without clearly indicating the much larger wage effects of occupation-average productivity growth.[2]

A statistician might say that the "control group is contaminated" because the productivity growth of the farmers is "spilling over" to barbers through labor markets. Our purpose here is not to discourage DiD analysis, even those with contaminated control groups, but rather to use price theory to help understand what DiD measures and how its findings can be applied to other settings.

III. Treatments and controls according to Marshall's laws

III.A. The derived demand for varieties in the industry model

To arrive at some formal results, suppose that we are interested in the own-price elasticity of the demand for F. Using the Δ operator to refer to the time derivative of the natural log, that elasticity ε_D could be estimated as $\Delta F/\Delta p$ if we were confident that F's price p was the only thing changing over time. The DiD approach is attractive when there are different varieties of F, which we call K and L, that experience different price changes Δr and Δw but are affected in the same way by changes in other factors denoted ΔA and $\Delta\theta$.

Although our emphasis is on the market incidence of treatment, familiar and succinct expressions are available if we describe market allocations as the choices of a representative consumer with preferences $u(AF(KB,L)/B,Z;\theta)$ over the quantities K, L, and Z, which is a composite quantity of all other goods. The aggregator F exhibits constant returns and has an elasticity of input substitution that we denote σ. We could dispense with the aggregator F and let K and L enter u separately, but this would make it more difficult to defend the "parallel trends" assumption that A and θ have no effect on the ratio K/L, as discussed further below.

In some applications, as with the farmers and barbers discussed at the beginning of this paper, K and L are aptly described as production factors. They could be labor in two different

[2] In a slightly different setting, DiD could show a negative relationship between occupation-specific productivity growth and occupation-specific wage growth even though economy-wide productivity increases wages. In such an example, the demand for, say, agricultural products is price inelastic. Productivity growth must therefore reduce agricultural employment. With imperfectly mobile labor in the short run, farm wages fall. Indeed, this is the economics storyline in Steinbeck's *The Grapes of Wrath* (1939). See also this chapter's Section V.

states, as in the minimum wage literature. They could be capital in two different industries. In other applications, K and L might represent distinct retail products, firms in the same industry that differ by size or location, or different sectors of the economy as in (Jaffe, et al. 2019, Chapter 17). The model is flexible in accommodating these cases. For the following analysis to be most applicable, the aggregate F of K and L should be of economic interest.

The scalars A and B are productivity parameters. A is F-input neutral, whereas B is biased toward K. θ is a preference parameter that shifts the marginal rate of substitution between the two arguments of u. θ may also be a way to model deviations from the representative consumer that do not violate parallel trends. These common shocks could be at the industry level if, for example, K and L were employment within an industry but having different locations. A and θ could be at the state level if K and L refer to employment in two different industries within the same state.

The consumer's budget constraint is $Z + rK + wL = M$, the left-hand-side of which can be written as $Z + PY$, where P is the marginal and average cost of the first argument Y of u. The parameter of interest $\varepsilon_D < 0$ is the price elasticity of the Marshallian demand for Y, which we denote as $D(P,M;\theta)$. The equilibrium can be described as:

$$D(P, M; \theta) = Y = \frac{A}{B} F(L, KB) \tag{1}$$

$$PY = wL + rK \tag{2}$$

$$K = \frac{\partial C(w, r, Y; A, B)}{\partial r} \;\; and \;\; L = \frac{\partial C(w, r, Y; A, B)}{\partial w} \tag{3}$$

where $C()$ is the minimum cost $rK + wL$ of "producing" Y.

Because the input-demand functions (3) are homogeneous of degree zero in prices and proportional to output, $C(w,r,Y;A,B)$ can be written as $C(Bw/r,1,1;1,1)rY/A$. Many of our results follow merely from totally differentiating this expression of the input demands and expressing them in terms of elasticities and shares:[3]

$$\Delta K = \Delta\left(\frac{Y}{A}\right) + s_L \sigma \Delta\left(\frac{w}{r}\right) + s_L \sigma \Delta B \tag{4}$$

$$\Delta L = \Delta\left(\frac{Y}{A}\right) - (1 - s_L)\sigma \Delta\left(\frac{w}{r}\right) + (1 - \sigma + s_L \sigma)\Delta B \tag{5}$$

[3] In doing so, we use the fact the elasticity of substitution in F is defined as $\sigma \equiv \frac{C}{(\partial C/\partial r)(\partial C/\partial w)} \frac{\partial^2 C}{\partial r \partial w}$. Because F exhibits constant returns, σ is independent of A and varies only with the ratio KB/L. Our appendix shows the full proof. For analysis of a composite good that is homothetic but not homogeneous in the inputs, our F can be interpreted as a monotone transformation of the composite that is homogeneous. In this case, the elasticities of demand for the composite would be rescaled versions of the corresponding elasticities of $D()$.

where s_L is L's expenditure share $wL/(rK+wL)$ and σ is the elasticity of substitution in F. None of our analytical results require that elasticities or shares are constant even though our prose may refer to them as "parameters."[4]

Various applications differ according to what is assumed about Δr, Δw, ΔK, ΔL, ΔA, and ΔB, implicitly within the constraints of (4) and (5). In all cases, ΔY is known as "a scale effect," having three key properties. First, ΔY is common to ΔK and ΔL. Second, ΔY is endogenous to "treatments," in a way that we make precise in what follows. Third, depending on the application, ΔY also depends on other factors such as M and θ that may not be part of the treatment. The first property means that the scale effect cannot be discovered merely by comparing ΔK to ΔL or Δr to Δw. Yet, the second property means that the scale effect is often of substantive interest. On the other hand, the third property reflects advantages to K/L and r/w comparisons in that they difference out the parts of the scale effect that are not causally linked to the treatment.

III.B. DiD estimators: Quantity outcomes

Take the case when a treatment works through prices – as with an excise or income tax, or certain regulations – and the outcomes are quantities. In that case Δr could be considered as the treatment and K as the treated. Equation (4) therefore describes the treated while equation (5) describes controls. $(1-s_L)$ is the expenditure share of the treated.[5]

An effect of other factors that is common to treatment and control groups is known in econometrics as "parallel trends." That is, if untreated, the treatment group would follow the same trend as the control group. The first terms in each of (4) and (5) show the three factors that are consistent with parallel trends: the input-neutral productivity change and anything with effects that operate solely through scale, such as income effects or taste changes.

The parameter of interest is the effect on ΔK, ΔL, and ΔF of treating the entire market ($\Delta r = \Delta w = 1$) with income and the other factors A, B, and θ held constant. We may be interested in the effect of an industry-wide cost change, but observe an instance when only one firm is treated. In other words, we are interested in the hypothetical of what would happen if all varieties were treated with the same price change. This hypothetical is a scale effect, which we show later to be summarized by the demand parameter ε_D. In contrast, the DiD estimator for the quantity effect of a price treatment is (6):

$$\frac{\Delta K - \Delta L}{\Delta r - \Delta w} = -\sigma + (\sigma - 1)\frac{\Delta B}{\Delta r - \Delta w} \tag{6}$$

[4] The shares and elasticities represent the values applicable to the point $\{w,r,Y,A,B\}$ where expression (3) has been totally differentiated.

[5] In other cases, we are also interested in quantity outcomes but the treatment is a productivity change rather than a price change. The same equations (4) and (5) are used for that purpose, as we do in Section V.

This estimator allows the "controls" to experience a price effect too ($\Delta w \neq 0$), as long as prices change more for the treated than for the controls. $\Delta w = 0$ is a special case, whereas (6) also describes those studies in the DiD spirit that have groups that are treated to different degrees.

Even without the biased-productivity term ΔB, the DiD estimator reveals substitution in F rather than the substitution toward other goods that is of primary interest. In Marshall's terms, DiD has given us the input-substitution effect $-\sigma$ when we want the scale effect ε_D. The equation (6) for the DiD estimator shows that the result would be $-\sigma$ even if the L variety were truly untreated in the sense that $\Delta w = 0$.

Many DiD studies with a time dimension begin by demonstrating parallel trends ($\Delta K \approx \Delta L$) in the pre-treatment data. We already noted how the A, θ, and M terms each contributes to parallel time series for K and L because the two varieties experience identical scale effects. The price terms might also coincide before the treatment occurs if $\Delta r = \Delta w$ during that time. But parallel pre-trends may also indicate that the two varieties are subject to the same market influences, in which case equilibrium spillover effects must be considered.

III.C. Spillover effects and the treatment on the treated

Equations (4) and (5) include the same scale term, which is potentially affected by the treatment. A full understanding of either the effect of the treatment Δr on the treated (TOT) or spillover effects therefore requires modeling the connection between scale and the treatment. Differential versions of the Y-demand and supply equations (1) and (2) provide such a model.

Specifically, the production part of equation (1), together with cost-minimization, requires that ΔY be the productivity terms plus the expenditure-share weighted average of ΔK and ΔL:

$$\Delta Y = \Delta A - s_L \Delta B + s_L \Delta L + (1 - s_L)\Delta K \tag{7}$$

This result and the pricing equation (2) require that ΔP be the same productivity terms plus the expenditure-share weighted average of Δr and Δw:

$$\Delta P = s_L \Delta w + (1 - s_L)\Delta r - (\Delta A - s_L \Delta B) \tag{8}$$

That is, equation (8) is a differential version of Y's marginal cost. The differential version of the demand part of equation (1) is:

$$\Delta Y = \varepsilon_D \Delta P + \eta_M \Delta M + \eta_\theta \Delta \theta \tag{9}$$

where η_M and η_θ denote the income- and taste-elasticity of the demand function $D(P,M;\theta)$. This setup allows Y (and therefore K and L) to have any income elasticity, including zero or negative values.

The reduced form of the supply-demand system (8)-(9) allows us to replace the scale terms in the conditional factor demands (4) and (5) to arrive at unconditional factor demands. These are

the same factor demand expressions that are the basis for the well-known Marshall's Laws of Derived Demand:

$$\Delta K = \eta_M \Delta M + \eta_\theta \Delta\theta + s_L(\sigma + \varepsilon_D)\Delta B - (1 + \varepsilon_D)\Delta A + [(1 - s_L)\varepsilon_D - s_L\sigma]\Delta r + s_L(\varepsilon_D + \sigma)\Delta w \quad (10)$$

$$\Delta L = \eta_M \Delta M + \eta_\theta \Delta\theta + [1 - \sigma + s_L(\sigma + \varepsilon_D)]\Delta B - (1 + \varepsilon_D)\Delta A + (1 - s_L)(\varepsilon_D + \sigma)\Delta r + [s_L\varepsilon_D - (1 - s_L)\sigma]\Delta w \quad (11)$$

where for conciseness we define the net demand shift $\Delta Z = \eta_M \Delta M + \eta_\theta \Delta\theta - (1+\varepsilon_D)\Delta A$.

As shown by Marshall (1895), any input's own-price elasticity is a weighted average of a scale effect ε_D and an input-substitution effect $-\sigma$ with the input-expenditure shares as weights.[6] Each cross-price elasticity is the sum of the two effects multiplied by the expenditure share of the input with the price change. Each of the ΔM, $\Delta\theta$, and ΔA terms is a product of an elasticity that is common to treatments and controls and a common change in one of the determinants of the demand for Y.

In applications where Δr is the treatment, the Δr coefficient in (10) is the TOT = $[(1-s_L)\varepsilon_D - s_L\sigma]$. The Δr term in (11) is the spillover effect of the treatment on the controls. Both price effects are decomposed into scale- and substitution-effect components, with their quantitative importance related to the share $(1-s_L)$ of the market that is treated. Equation (10) also provides the proof that the effect of a price treatment for the entire market ($\Delta r = \Delta w > 0 = \Delta B = \Delta M = \Delta\theta = \Delta A$) is ε_D, in contrast to the DiD estimator of $-\sigma$ shown in equation (6).

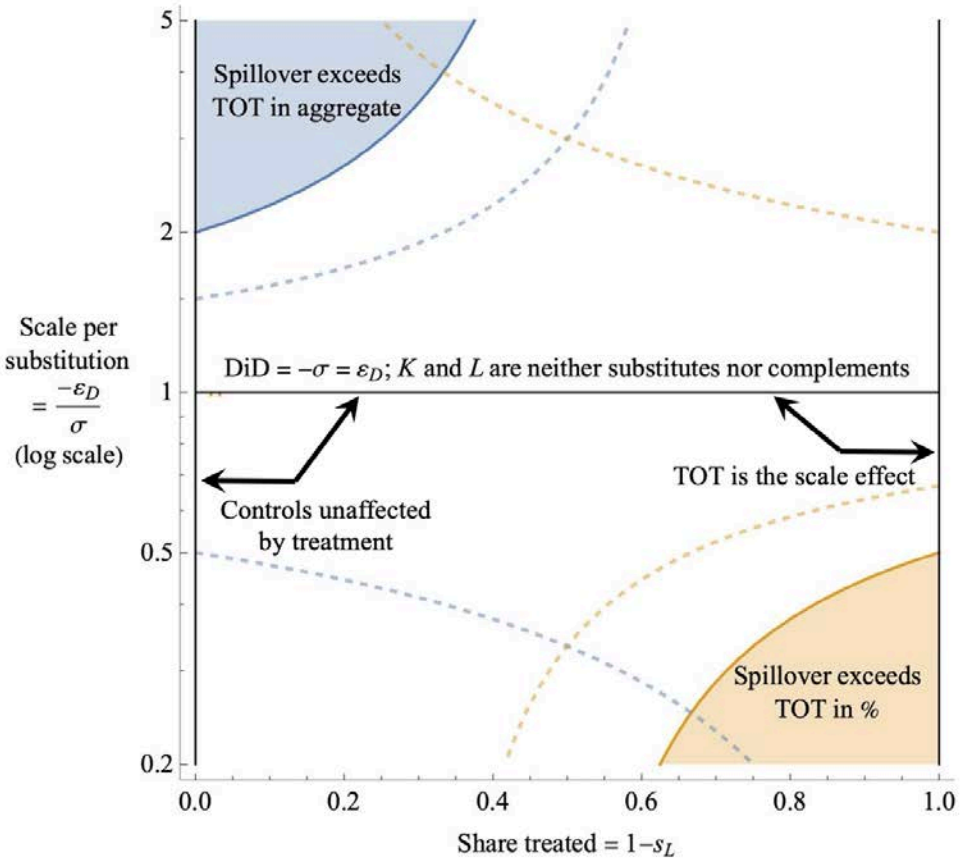
III.D. The treatment share and the equilibrium spillover effect

Here we show how the share $(1-s_L)$ of the market that is treated affects the interpretation of DiD estimates. Small treatment shares have the advantage of spillover effects that are small compared to the effect of the treatment on the treated *when both are measured in percentage terms*. However, this comes with two disadvantages. One is that, surprisingly, the spillover effect is comparatively large in the aggregate. The other is that scale effects are obscured.

When $\Delta w = \Delta B = 0$, Figure 1 shows the various possibilities for the spillover effect (the Δr term in equation (11)) as compared to the effect of the treatment on the treated (TOT; the Δr term in equation (10)). The horizontal axis is the expenditure share $(1-s_L)$ of the treated. The vertical axis is the ratio of the magnitude $-\varepsilon_D$ of the scale elasticity to the input-substitution elasticity σ. The horizontal line shows all the possibilities satisfying $\varepsilon_D + \sigma = 0$, which means that the two varieties are neither substitutes nor complements.

[6] Equations such as (10) or (11) expressing Marshall's Laws of Derived Demand are distinct from the Slutsky equation in terms of the scale effect. The Slutsky equation holds constant input expenditure $wL + rK$ whereas Marshall's Laws are equilibrium comparative statics: expenditure increases or decreases as necessary to keep the output price P and quantity Y on the market demand curve $Y = D(P,M;\theta)$.

Figure 1. The share treated and the equilibrium spillover effect



Notes: TOT = Treatment on the Treated. The dashed curves show parameter values where the spillover effect magnitude is exactly half TOT, either in aggregate (dark) or percentages (light)

On the horizontal line, both cross-price terms in equations (10) and (11) are zero. A zero cross-price term in (11) means that the treatment does not spill over onto the controls. The other way that the spillover effect can be zero is on the vertical line at zero treatment share. However, the other cross term is not zero on that line except where it intersects the horizontal line. When the two varieties are either substitutes or complements, treating the controls could have an important effect on the treatment group precisely because the treatment group is relatively small. That cross term from equation (10) is part of the effect of treating the entire market.

In other words, having a treatment share close to zero helps solve the control-contamination problem (more on this below) but at the expense of increasingly weighting the TOT = $[(1-s_L)\varepsilon_D - s_L\sigma]$ toward the input-substitution effect σ rather than the scale effect ε_D that we want. In order for the TOT to approximate the scale effect, the share treated must be close to one, or that the scale and substitution effects essentially cancel as they do on the horizontal line.[7]

[7] σ itself may vary with s_L. Regardless, no value of s_L will intrinsically tie σ to the parameter of interest ε_D.

Figure 1 also shows three light-colored curves: one solid and two dashed. The solid curve represents those parameter combinations where the spillover effect, in percentage terms, has the same magnitude as the TOT. The area below that curve represents parameter combinations where the spillover effect has, in percentage terms, the greater magnitude. These parameters are on the bottom right of Figure 1 because the spillover effect is relatively large when a large share of the market is treated, especially when the substitution effect is large.

The effect of the treatment share ($1-s_L$) can be illustrated with a geographic example. Let K represent Canada and L the rest of the world (ROW). Canada has about one percent of world income and about 0.5 percent of world population. With a treatment share $1-s_L$ so near zero, a public policy implemented in Canada alone ($\Delta r > 0 = \Delta w$) is unlikely to have a noticeable effect on the ROW's economic or demographic statistics. Nevertheless, implementing the same policy throughout the ROW may well have a significant effect on Canada. If only 0.1 percent of the world's population decided to move to Canada, that would increase Canada's population by 20 percent. Comparing Canada's before-after to the before-after of an untreated but otherwise similar country shows the Δr effect but not the Δw effect that Canada would experience if the ROW were treated.

In percentage terms, the spillover effect can still be significant by comparison with the TOT, even if the TOT has the greater magnitude. The two light-colored and dashed curves show parameters for which the magnitude of the spillover effect is exactly half that of the TOT. Parameters above the upper dashed curve have a spillover effect that is more than half TOT.[8] As expected from the Canadian example, increasing the treatment share increases the spillover effect by a greater proportion than it changes the TOT. Even at a small treatment share, the magnitude of the spillover effect can be arbitrarily close to that of the TOT if the scale effect is large enough.

For some purposes, such as welfare analysis, the aggregate spillover effect is important. Figure 1's solid dark curve represents those parameter combinations where the spillover effect has the same magnitude as the TOT in aggregate. Surprisingly, reducing the treatment share reduces the magnitude of the aggregate TOT by a greater proportion than it reduces the aggregate spillover effect, if the spillover is reduced at all. The more significant aggregate spillover effects are therefore shown in Figure 1 in the lower-left corner and, especially, the upper-left corner.

The reason for this surprising result can be understood by considering the substitution and scale effects separately. Without any scale effect ($\varepsilon_D = 0$), the aggregate spillover effect would be equal to the aggregate TOT effect, albeit in the opposite direction, regardless of s_L. The equal and opposite aggregate effects is essentially the definition of a substitution effect. The scale effect, on the other hand, contributes equal *percentage* changes to both K and L. Decreasing K's share therefore reduces the scale component of the aggregate TOT while it increases the scale component of the spillover effect.

[8] The sign comparison of the TOT and spillover effect depends on whether the varieties are substitutes (opposite sign; bottom half of Figure 1) or complements (same sign; top half of Figure 1).

Algebraically, the necessary and sufficient condition describing the upper-left area in Figure 1 is:

$$\left[\frac{s_L}{1-s_L}\frac{(1-s_L)(\varepsilon_D+\sigma)}{(1-s_L)\varepsilon_D-s_L\sigma}\right]^2 > 1 \Leftrightarrow \left(s_L-\frac{1}{2}\right)\varepsilon_D < -s_L\sigma \tag{12}$$

where the term in square brackets has two fractions. The first fraction, which is the ratio of the control share to treatment share, is needed because we refer to aggregates. The second ratio has the spillover and direct effects as numerator and denominator, respectively, each expressed as elasticities. The simpler equivalent expression shows that the condition simultaneously requires the control group to be larger ($s_L > ½$) and K and L to be complements ($\varepsilon_D + \sigma < 0$). As the share treated becomes small, the inequality (12) reduces to $\varepsilon_D < -2\sigma$.

Suppose, for example, that $\varepsilon_D = -2$ and $\sigma = 1/2$. In aggregate, the direct effect of Δr is $(1-s_L)[3s_L/2 - 2]$ and the spillover effect is $-s_L(1-s_L)3/2$. The spillover effect has greater magnitude for any $s_L > 2/3$. In the limit as s_L approaches one, the spillover effect is three times the direct effect.

III.E. DiD estimators: Price outcomes

A third category of DiD studies features quantity or productivity treatments with price outcomes. For the union wage effect that we examine in Section VI, the quantity treatment comes from efforts by trade unions to reduce the supply of labor to the union sector with the intended effect of raising wages in that sector. Other studies have looked at the price effects of the sudden shutdown of a factory, perhaps by natural disaster or by regulation.[9]

Quantitative relationships between treatment and spillover effects on price outcomes may be examined by solving the system (10) and (11) for prices, as in (13) and (14):

$$\Delta r = \frac{1}{\varepsilon_D}\left[[(1-s_L)\sigma - s_L\varepsilon_D]\frac{\Delta K}{\sigma} + s_L(\varepsilon_D+\sigma)\frac{\Delta L-\Delta B}{\sigma} - \Delta Z\right] \tag{13}$$

$$\Delta w = \frac{1}{\varepsilon_D}\left[(1-s_L)(\varepsilon_D+\sigma)\frac{\Delta K}{\sigma} + [s_L\sigma - (1-s_L)\varepsilon_D]\frac{\Delta L-\Delta B}{\sigma} - \Delta Z\right] - \Delta B \tag{14}$$

$$\Delta P = \frac{1}{\varepsilon_D}\left[(1-s_L)\Delta K + s_L\frac{\Delta L-\Delta B}{\sigma} - \Delta Z\right] - \Delta A \tag{15}$$

where, as before, ΔZ denotes the net demand shift $\eta_M\Delta M + \eta_\theta\Delta\theta - (1+\varepsilon_D)\Delta A$. We add the analogous expression (15) for the price of Y, which can be subtracted from (13) and (14) as needed for "real wage" expressions.

[9] Hakim, Gupta and Ross (2017) examines effects of regulator-required factory closures on retail prices in the market for generic drugs.

Now we can return to the farmers from the beginning of this paper. They were "treated" with productivity growth that was not experienced by barbers. The parameter of interest is the real-wage effect of productivity growth in all occupations, represented as $\Delta A > \Delta B = 0$. From (13)-(15) we have:

$$(1 - s_L)\Delta\left(\frac{r}{P}\right) + s_L\Delta\left(\frac{w}{P}\right) = \Delta A \qquad (16)$$

That is, neutral productivity growth increases real wages by the same proportion, although the allocation of the wage increase between r and w can be uneven.[10] The average real-wage effect is independent of the elasticity of input substitution σ. In contrast, equation (6) shows that the effect of biased productivity growth ($\Delta B > 0$) on the DiD $\Delta r - \Delta w$ depends only on σ and the effect of relative price changes on relative input supplies. For example, in the price theory oral tradition, $\Delta r - \Delta w = 0$ because the two inputs are perfect substitutes on the supply side. In other words, market-wide productivity growth increases wages generally even though biased productivity growth does not affect relative wages. See also Section V.

IV. Complementing DiD with price theory

The straightforward case for generalizing a DiD estimate is when we are interested in the effect on K of $\Delta r > 0 = \Delta w$ rather than the effect of a hypothetical aggregate treatment. Price theory tells us that the coefficient is $[(1-s_L)\varepsilon_D - s_L\sigma]$, which is similar to the DiD estimate when the share treated $(1-s_L)$ is close to zero. For example, we could use a DiD estimate of Canada's policy experience as an estimate of what would happen to another small and otherwise similar country that might adopt the same policy because both of them would be $-\sigma\Delta r$.

Even when the parameter of interest is the effect ε_D of treating the entire market, price theory shows how DiD can be part of obtaining a reliable estimate. Two instances follow.

IV.A. DiD indicates the correction required for uneven treatments

The idea that the L variety is "contaminated" by K's price change is represented by the presence of Δr in equation (11) to the extent that $\varepsilon_D \neq -\sigma$. Nevertheless, price theory shows how the DiD estimator can be useful in recovering ε_D. To see this consider rewriting (10) as (16):

$$\Delta K = \eta_M\Delta M + \eta_\theta\Delta\theta - (1 + \varepsilon_D)\Delta A + \varepsilon_D\frac{\Delta r + \Delta w}{2} + \left[s_L\sigma + \left(s_L - \frac{1}{2}\right)\varepsilon_D\right](\Delta w - \Delta r) \qquad (17)$$

[10] Depending on the supply conditions for K and L, which are unrestricted by our basic model (1)-(3), relative wage changes may be required to motivate proportional increases in these two quantities. However, nonhomothetic supply conditions would make it more difficult to defend the parallel trends assumption.

In words, equation (16) separates the price effects on K into two terms: (i) an average price-change term whose coefficient is the parameter of interest ε_D, and (ii) a correction term accounting for inequality of the price changes. Calculating the correction term is facilitated by having an estimate of σ, which DiD can provide as in equation (6).[11] Although DiD misses much, if not all, of the scale effect, it can be a tool for estimating the scale effect by providing quantitative information about the input-substitution effect.

IV.B. Outside- and within-market control groups

Because the spillover effects occur through the market, another approach might be to select an additional control group satisfying three properties: (i) from a market different from the treatment group, (ii) having $\Delta w = 0$, and (iii) nonetheless experiencing similar trends ΔA, $\Delta\theta$ and ΔM for the other factors. The within-market DiD estimator (6) would yield $-\sigma$, but the cross-market DiD estimator would be:

$$\frac{\Delta K - \Delta L}{\Delta r - \Delta w} = \frac{[(1 - s_L)\varepsilon_D - s_L\sigma]\Delta r}{\Delta r - 0} = (1 - s_L)\varepsilon_D - s_L\sigma \tag{18}$$

With data on the treated-group share ($1-s_L$) and the two DiD estimates, the scale effect ε_D could be recovered. Specifically, the scale effect is the weighted average of the two DiD estimates, with negative weight on (6).[12] In effect, the two DiD estimates permit extrapolating to a hypothetical situation in which the entire market is treated.

V. Difference-in-Differences may indicate the wrong sign

Now suppose that the question of interest is the effect of the productivity parameter A on F. For example, we want to know how productivity affects an industry's overall employment of the factors of production. The DiD method gets our attention because we suspect that there are price changes largely common to K and L that are unmeasured but changing coincident with our measures of A. We look for a situation where, like the farmers from the beginning of this paper, K is more "treated" with productivity growth than L is. Formally, we have $\Delta r = \Delta w$ while $\Delta B > 0$. Now (10) and (11) become:

$$\Delta K = \varepsilon_D\Delta r + \eta_M\Delta M + \eta_\theta\Delta\theta - (\varepsilon_D + 1)\Delta A + s_L(\varepsilon_D + \sigma)\Delta B \tag{19}$$

$$\Delta L = \varepsilon_D\Delta r + \eta_M\Delta M + \eta_\theta\Delta\theta - (\varepsilon_D + 1)\Delta A + [1 - \sigma + s_L(\varepsilon_D + \sigma)]\Delta B \tag{20}$$

From (18) and (19) we see that the parameter of interest is $-(\varepsilon_D+1)$ because that is the rate at which A increases both inputs. The effect is positive if and only if industry demand is price elastic. The DiD estimate of the effect of productivity on factor quantity is:

---

[11] ε_D can be calculated from ΔK, Δr, Δw, ΔA, s_L, and σ: $\varepsilon_D = \frac{\Delta K + s_L\sigma(\Delta r - \Delta w) + \Delta A - \eta_M\Delta M + \eta_\theta\Delta\theta}{s_L\Delta w + (1-s_L)\Delta r - \Delta A}$.

[12] The weight on (6) is $-s_L/(1-s_L)$.

$$\frac{\Delta K - \Delta L}{\Delta B} = \sigma - 1 \tag{21}$$

In other words, the DiD estimate is positive if and only if *K* and *L* are sufficiently good substitutes in *F*. The DiD estimate could be negative (low substitution in *F*) while the parameter of interest is negative (good substitution with the outside good), or vice versa. In either of these possibilities, DiD has the wrong sign.

Consider again this paper's story of farmers and barbers. With inelastic demand for farm products, productivity growth in farming must shift labor out of farming (*K*) and into other occupations (*L*). According to equation (20), $\sigma < 1$ and the DiD estimate would be negative. Nevertheless, the aggregate amount of labor could increase; that depends on the sign of ε_D+1 in our notation.

With (18)-(20) focused on effects of productivity growth, we assume $\Delta r = \Delta w$ in order to justify the parallel trends assumption. With the farmers and barbers, $\Delta r = \Delta w$ is not an assumption but rather a result of the more fundamental assumptions about occupational choice. Either way, the DiD estimate of the effect of productivity on factor prices is zero: $(\Delta r - \Delta w)/\Delta B$. The aggregate effect of interest is that neutral productivity growth ΔA increases r/P and w/P by the same proportion.

Even in the context of the price-quantity relationships (10) and (11), the DiD estimator could have the opposite sign as the effect of the treatment on *F*. That must occur when *K* is an inferior input, which we rule out in this paper by assuming that *F* is homothetic. Increasing the price of the inferior input leads to substitution to the luxury input, which is no surprise from the DiD perspective (6). The surprise is that the same input-price increase reduces the marginal cost of *F*, which is a scale effect *increasing F*. Mulligan (2022) applies this framework to understanding the decline in life expectancy after 2010 by interpreting *K* as the quantity of prescription opioids, *L* as the quantity of illicitly-manufactured opioids, and *F* as opioid mortality. This helps explain why various empirical studies conclude that opioid mortality increased due to regulations increasing the full price of prescriptions.

VI. Further examples of difference-in-differences in the marketplace

VI.A. The union wage effect

The union wage effect has been studied with the DiD method, with wages as the outcome. In our notation, the outcome for the treated is *r*, which is compared to the factor rental-rate *w* for the non-union "controls." Here we interpret the unionization "treatment" as a restriction on the supply *K* of labor in the unionized sector to raise wages *r* in that sector. Licensing requirements are examples. For simplicity, we assume that workers unable to gain employment in the unionized

sector are employed in the nonunion sector, a supply condition that is formally represented as $dK + dL = 0$ (Rees 1989). With this restriction, equations (10) and (11) become (21) and (22):

$$\Delta\frac{r}{p} = \Delta A - \frac{s_L}{q_L}\frac{\Delta K}{\sigma} \tag{22}$$

$$\Delta\frac{w}{p} = \Delta A + \frac{1 - s_L}{q_L}\frac{\Delta K}{\sigma} \tag{23}$$

where q_L denotes the nonunion sector's quantity share $L/(K+L)$, which is equal to its factor share s_L absent the union-supply restriction. The more that supply is restricted, the more than q_L exceeds s_L. Note that (21) and (22) imply that unionization increases wages in the union sector while reducing wages elsewhere with magnitudes (that differ when $s_L \neq 1/2$) governed by s_L, q_L, and σ.

The DiD estimator is the union-nonunion wage gap:

$$\Delta\frac{r}{p} - \Delta\frac{w}{p} = -\frac{1}{q_L}\frac{\Delta K}{\sigma} \tag{24}$$

The DiD estimator (23) is different from equation (21), which is the treatment effect of unionization on the $1-q_L$ who are unionized. Their difference is equation (22), whose final term quantifies the "contamination of" (or treatment spillover onto) the controls. The term would be near zero if the nonunion share were close to one. However, studies of union wages often include markets with sizeable union sectors.[13] In such cases, much of the union-nonunion wage gap may reflect a reduction in nonunion wages rather than an increase in union wages.[14] Even if the union sector were small, the effects on non-union wages of supply constraints in the union sector could exceed the TOT in aggregate because of the relatively large number of workers in the nonunion sector.

Both equation (21)'s final term and the DiD estimator (23) and are different from (22)'s final term times −1, which would be the effect of unionizing the remaining q_L of the workforce by restricting that supply by the same proportion.[15] Particularly when the unionization rate is low, the effect on the wages of erstwhile nonunion workers of extending union status to them would be much less than indicated by the union-nonunion wage gap (23). These are further examples of

[13] Referring to the year 1977, Freeman and Medoff (1984, Table 2-1) estimate that 30 percent of blue-collar workers were unionized, with a unionization rate of 61 percent in "Transportation, communication, and other public utilities."

[14] Another interesting "spillover" effect of unionization is the effect on wages in non-unionized firms in the same sector. Studies such as Rosen (1969) suggest that those wages are increased due to a "union threat" effect.

[15] An equi-proportional reduction in the supply of both K and L would have no effect on either r/p or w/p. Therefore going from restricting K only to restricting both K and L would reverse the ΔK term in the equation (22) for $\Delta(w/p)$. The effect on Δw is found by adding $\Delta(w/p)$ to Δp, which is purely a scale effect term.

how, in market settings, the DiD estimator differs from parameters that are potentially more interesting.[16]

VI.B. Models with time and region fixed effects

Without price theory as a guide, difference-in-differences estimates can easily be misinterpreted in geographical contexts. One case is an early set of studies attempting to detect imperfect competition in cigarette manufacturing in the form of "over-shifting" cigarette excise taxes (Sumner 1981). Over-shifting means that a causal effect of a $1 per pack tax is to increase the retail price of cigarettes by more than $1 per pack, whereas "one-for-one passthrough" refers to a dollar-for-dollar correspondence between excise taxes and retail prices. These studies were executed with essentially a difference-in-differences framework by comparing states with large tax increases to states with little or no increase. They found nearly one-for-one pass through, but did not consider whether the retail prices in the control states were increased by the tax rates in the treatment states.[17] If the control states were affected in this way, nationwide increases in excise taxes would be over-shifted even though the state DiD shows one-for-one pass through.

Another example is related to (Jaffe, et al. 2019, Chapter 17), which concludes that business taxes reduce wages in the long run because the taxes reduce productivity. Nevertheless, an increase in business taxes in a particular locality may not reduce wages in that locality relative to the rest of the nation because workers have a choice of where to live and work. In effect, the wage in any locality is influenced by business taxes throughout the country, or even throughout the world. By failing to pick this up, a DiD approach might not show any wage effect of business taxes for much the same reason discussed at the beginning of this chapter in the occupational context.

If geographic differences in business taxes result in little or no geographic differences in wages, they might result in especially large geographic differences in employment. This is another case in which the geographic-specific effect is different from the aggregate effect, but this time with the former effect being greater.

Another policy question is the employment effect of public projects such building a sports stadium or hosting a major event such as the Olympics. Early studies used something like a DiD approach and found a "multiplier": that total employment in the vicinity of the stadium increased more than the number directly employed by the sports enterprise (Wanhill 1983, Johnson, Obermiller and Radtke 1989). One reason for this is that complementary businesses were opened nearby, such as restaurants, lodging, and parking. But later studies found that most, if not all, of the additional employment was pulled in from other localities (Dwyer and Forsyth 2009).

---

[16] Studies of the union wage effect often do not draw distinctions between DiD, TOT, etc. For example, Freeman (1984) refers to "the effect of unionism" and the "true impact of unionism."

[17] Suppose, for example, cigarette manufacturers set one nationwide wholesale price because of concerns that regional wholesale price inequality would result in unauthorized wholesale orders and shipments in the low-price regions on behalf of the high-price regions. Such manufacturers would respond to an increase in one state's excise rate by adjusting their nationwide wholesale price, and through that mechanism indirectly adjust retail prices throughout the nation. Later studies acknowledged this market mechanism's effect on state differences (Keeler, et al. 1996, Evans, Ringel and Stech 1999, Adhikari 2004); see also Tennant (1950). Harris (1987) emphasizes the results of a federal tax change.

Development economics includes experiments that encourage healthcare providers in treatment villages to supply more healthcare. Others incentivize more instructional effort by teachers in the treatment villages. Such experiments can be analogous to the sports-stadium studies. Namely, through factor markets the experiment reallocates resources from control villages to treatment villages. The per-capita effect of treating all villages would be different unless resources are moved with equal easy (or difficulty) between villages as from outside the village economy as a whole. In our notation, that condition is $\varepsilon_D + \sigma = 0$.

VII. Summary and conclusions

Markets are ubiquitous. Consumers and businesses do not live or work in isolation, even approximately so. Perhaps one reaction among those engaged in measurement is to actively attempt to isolate members of the treatment group. Clinical drug trials, for example, do try to prevent trial participants from trading with each other, that is, sharing or exchanging the treatments with others. Some clinical trials even discourage participants from communicating specifics about their trial experiences to prevent (what the investigators view as) potential bias or cross-contamination of results.

We take a different approach in this paper, which is to acknowledge trade and keep it at the center of the analysis. The prototype supply and demand model can by itself account for treatments and controls in a market setting, with Marshall's Laws of Derived Demand showing how difference-in-differences methods tend to reveal substitution effects rather than scale effects. It's crucial to measure the extent of actual and hypothetical "treatments" in order to accurately interpret, and generalize from, difference-in-differences estimates.

What econometricians sometimes call "spillover" effects are not well described as externalities – missing markets – because markets also transmit treatment effects to the untreated through prices. Analogizing spillover effects with externalities may give the wrong impression that such effects are rare or beyond basic economic training.

At least in the market setting, per capita spillover effects tend to decrease as the size of the treatment group goes to zero, but so does the aggregate treatment effect. Small-scale treatments thereby come with two disadvantages. One is that scale effects are especially obscured by substitution effects. Second, and surprisingly, the spillover effect is comparatively large in the aggregate.

References

Adhikari, Deergha Raj. 2004. "Measuring market power of the US cigarette industry." *Applied Economics Letters* 11: 957–959.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2008. *Mostly harmless econometrics: An empiricist's companion.* Princeton, NJ: Princeton University Press.

Athey, Susan, and Guido W. Imbens. 2017. "The state of applied econometrics: Causality and policy evaluation." *Journal of Economic Perspectives* 31: 3–32.

Borusyak, Kirill, Peter Hull, and Xavier Jaravel. 2022. "Quasi-experimental shift-share research designs." *The Review of Economic Studies* 89: 181–213.

Crépon, Bruno, Esther Duflo, Marc Gurgand, Roland Rathelot, and Philippe Zamora. 2013. "Do labor market policies have displacement effects? Evidence from a clustered randomized experiment." *The quarterly journal of economics* 128: 531–580.

De Chaisemartin, Clément, and Xavier d'Haultfoeuille. 2020. "Two-way fixed effects estimators with heterogeneous treatment effects." *American Economic Review* 110: 2964–2996.

Dwyer, Larry, and Peter Forsyth. 2009. "Public sector support for special events." *Eastern Economic Journal* 35: 481–499.

Egger, Dennis, Johannes Haushofer, Edward Miguel, Paul Niehaus, and Michael Walker. 2022. "General equilibrium effects of cash transfers: experimental evidence from Kenya." *Econometrica* 90: 2603–2643.

Evans, William N., Jeanne S. Ringel, and Diana Stech. 1999. "Tobacco taxes and public policy to discourage smoking." *Tax policy and the economy* 13: 1–55.

Freeman, Richard B. 1984. "Longitudinal Analyses of the Effects of Trade Unions." *Journal of Labor Economics* 2 (1): 1-26.

Freeman, Richard B., and James L. Medoff. 1984. *What do unions do?* New York: Basic Books.

Glaeser, Edward L., and Joshua D. Gottlieb. 2009. "The wealth of cities: Agglomeration economies and spatial equilibrium in the United States." *Journal of economic literature* 47: 983–1028.

Goldsmith-Pinkham, Paul, Isaac Sorkin, and Henry Swift. 2020. "Bartik instruments: What, when, why, and how." *American Economic Review* 110: 2586–2624.

Hakim, Aaron, Ravi Gupta, and Joseph S. Ross. 2017. "High costs of FDA approval for formerly unapproved marketed drugs." *JAMA* 318: 2181–2182.

Harris, Jeffrey E. 1987. "The 1983 increase in the federal cigarette excise tax." *Tax policy and the economy* 1: 87–111.

Hong, Guanglei, and Stephen W. Raudenbush. 2006. "Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data." *Journal of the American Statistical Association* 101: 901–910.

Huber, Martin. 2023. *Causal Analysis.* Cambridge: The MIT Press.

Hudgens, Michael G., and M. Elizabeth Halloran. 2008. "Toward causal inference with interference." *Journal of the American Statistical Association* 103: 832–842.

Jacobs, Jane. 1969. "Strategies for helping cities." *The American Economic Review* 59: 652–656.

Jaffe, Sonia, Robert Minton, Casey B. Mulligan, and Kevin M. Murphy. 2019. *Chicago Price Theory.* Princeton University Press (ChicagoPriceTheory.com).

Johnson, Rebecca L., Fred Obermiller, and Hans Radtke. 1989. "The economic impact of tourism sales." *Journal of Leisure Research* 21: 140–154.

Keeler, Theodore E., Teh-wei Hu, Paul G. Barnett, Willard G. Manning, and Hai-Yen Sung. 1996. "Do cigarette producers price-discriminate by state? An empirical analysis of local cigarette pricing and taxation." *Journal of health economics* 15: 499–512.
Krueger, Anne O. 1991. "Report of the commission on graduate education in economics." *Journal of Economic Literature* 29: 1035–1053.
Manski, Charles F. 1993. "Identification of endogenous social effects: The reflection problem." *The review of economic studies* 60: 531–542.
Marshall, Alfred. 1895. *Principles of Economics.* London: MacMillan and Co.
Miguel, Edward, and Michael Kremer. 2004. "Worms: identifying impacts on education and health in the presence of treatment externalities." *Econometrica* 72: 159–217.
Monte, Ferdinando, Stephen J. Redding, and Esteban Rossi-Hansberg. 2018. "Commuting, migration, and local employment elasticities." *American Economic Review* 108: 3855–3890.
Mulligan, Casey B. 2022. "Prices and Policies in Opioid Markets." *NBER working paper* (26812).
Rees, Albert. 1989. *The economics of trade unions.* University of Chicago Press.
Rosen, Sherwin. 1969. "Trade union power, threat effects and the extent of organization." *The Review of Economic Studies* 36: 185–196.
Sobel, Michael E. 2006. "What do randomized studies of housing mobility demonstrate? Causal inference in the face of interference." *Journal of the American Statistical Association* 101: 1398–1407.
Steinbeck, John. 1939. *The grapes of wrath.*
Sumner, Daniel A. 1981. "Measurement of monopoly behavior: an application to the cigarette industry." *Journal of Political Economy* 89: 1010–1019.
Tennant, Richard B. 1950. "The American cigarette industry: a study in economic analysis and public policy." *(No Title).*
U.S. Department of Agriculture, National Institute of Food and Agriculture. 2014. *About us: extension.* March 28. Accessed May 16, 2014. https://web.archive.org/web/20130422034544/http://www.csrees.usda.gov/qlinks/extension.html.
Wanhill, Stephen R. C. 1983. "Measuring the economic impact of tourism." *The Service Industries Journal* 3: 9–20.