Difference-in-Differences in the Marketplace[*]

by Robert Minton and Casey B. Mulligan[†]

April 2024

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 provides a simple statistical framework highlighting that controls are indirectly affected by the treatment through the market. Moreover, even the effect of the treatment on the treated reveals only part of the consequence for the treated of treating the entire market. When combined with economic theory, the statistical assumption of parallel trends leads to a new application of Marshall's Laws of Derived Demand. Emphasizing a close connection between treatment effects and the scale and substitution effects featured in price theory, Marshall's Laws show how difference-in-differences can diverge – both in magnitude and direction – from the causal effects of treating all market participants.

[*] 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 Jim Heckman, Allen Sanderson, Alex Torgovitsky, Giuseppe Forte, Josh Gross, João Pugliese, and Alex Tordjman.

[†] Affiliations and contact: Federal Reserve Board, Robert.j.minton@frb.gov and University of Chicago, c-mulligan@uchicago.edu, respectively.

I. Introduction

Markets are likened to an invisible hand. Among other things, the hand coordinates individual choices with the decisions of all other participants on the same side of the market, regardless of whether they directly interact. The invisible hand appears to contradict 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.

Section II of this paper illustrates the invisible hand 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. It shows that market spillovers can be the dominant factor determining outcomes, in at least some important contexts. Section III provides a simple statistical framework that allows for market spillovers and incorporates two concepts of "parallel trends" for treatments and controls. When treated and control observations are in the same market, the controls are indirectly affected by the treatment. Even without choice-theoretic restrictions, the framework indicates quantitative relationships between difference-in-differences estimates and meaningful counterfactuals.

Another role of prices in equilibrium models is to equalize quantities supplied and demanded. This function is left implicit in our analysis because it is already familiar in econometrics, particularly regarding the simultaneous feedback between supply and demand schedules. We focus on relating restrictions from demand theory to parallel trends assumptions. In the choice framework, parallel trends require the treatment and control outcomes to be weakly separable in utility, production, or cost from all other outcomes. This leads to Marshall's Laws of Derived Demand as a source of precise price-theoretic interpretations of the direct and spillover effects of a treatment.

A DiD estimator measures the degree of substitution between treatments and controls, regardless of the fraction of the market that is treated or the magnitude of market spillovers. In contrast, the effect of treating the entire market is a "scale effect," which is the price-theoretic term for the degree of substitution with goods outside the market where treated and controls participate. The effect of the treatment on the treated (ToT) is a weighted average of the scale effect and the DiD, whereas the market spillovers are proportional to their difference.

Our price-theoretic analysis classifies treatments and outcomes into three distinct categories. For treatments, they are prices, quantities, and productivity. For outcomes, the categories are quantities, prices, and expenditures. In some cases, the law of demand requires that the scale effect and DiD have the same sign. However, with productivity treatments or expenditure outcomes, the two can have opposite signs. In other words, DiD can have the opposite sign of the scale effect, and potentially of the ToT, for purely economic reasons.

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.

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 relate to broader treatment effects on the entire market. Our approach focuses on spillovers mediated by market forces as opposed to 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. See also Banzhaf (2021). 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. Cautioning against “inattention to the market consequences of the [programs evaluated],” Heckman, Lochner, and Taber (1999) provide an equilibrium model for evaluating both behavior and welfare effects of tuition subsidies and other public policies. Heckman, LaLonde, and Smith (1999) conclude that “the costs of ignoring indirect [equilibrium] effects may be substantial.” 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 in different states.

Our contribution to this research space, into which we have provided only a small glimpse, is a versatile and concise 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.

[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.”

Munro et al.'s (2021) observation that "the interference pattern produced by marketplace price effects is dense and simultaneously affects all units, so cluster- or sparsity-based methods are not applicable" is consistent with the price theoretic view of market equilibrium taken in this paper. They refer to a "Global Treatment Effect (GTE)" which we link to the scale effect from price theory. Both our paper and theirs treat this as "a meaningful policy-relevant counterfactual of treating all individuals in the [market] compared to treating no individuals in the [market]." Although they examine an experimental setting, their "average direct effect" is closely analogous to what we call the "difference-in-differences" estimator.

Drawing from market equilibrium analysis, we emphasize that the treated and untreated experience scale and substitution effects in different combinations.[2] This analytical approach has parallels with Heckman and Vitlaycil's (2005) expression of estimators as combinations of "marginal treatment effects," each of which refers to a specific type of individual. To focus on the price theoretic components, this paper considers only limited heterogeneity, namely treated versus untreated and in-market versus out-of-market. It emphasizes market connections, with counterfactual treatment regimes understood as additional distinct combinations of scale and substitution effects.

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 real-wage effect, if any. Conversely, if

[2] Much quantitative work in both micro- and macro-economics treats scale and substitution parameters as constants.

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.[3]

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. A vector representation of market spillovers

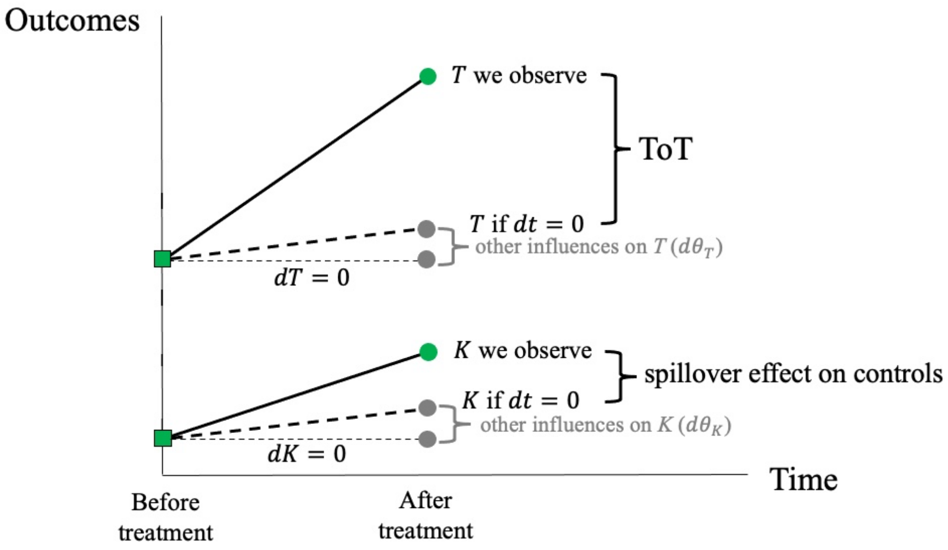
To begin the formal analysis, we consider a population of agents that are designated either as treatments or controls. Their population shares are denoted τ and $1-\tau$, respectively. Their outcomes are denoted T and K, respectively. The treatment, which directly affects the treated but not the controls, is denoted t. We let k denote a comparable shock that directly affects the controls but not the treated. The mappings from the two treatments to outcomes are denoted $T(t,k;\theta_T)$ and $K(t,k;\theta_K)$, where θ_T and θ_K denote other factors that influence outcomes for treated and controls.

Figure 1 illustrates the effects of a treatment $dt = 1$ in the time dimension. The familiar parts of the diagram are that (i) other factors influence both the T and K outcomes over time and (ii) the effect of the treatment on the treated (ToT) is the difference between the final outcome for the treated and what that outcome would be without treatment. Given our emphasis on market connections between T and K, our Figure 1 also allows for an effect of $dt = 1$ on the outcome for the controls.

[3] 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 paper's Section V.

Figure 1. Treatments and Market Spillovers

in the time dimension



Notes: For the purposes of Figure 1, the controls are untreated ($dk = 0$) but in the same market. Observations reflect a treatment of $dt = 1$ directly affecting the treatment group.

To focus on equilibrium interpretations of DiD estimates, we maintain the parallel trends assumption that $d\theta_T = d\theta_K$ and have the same marginal effects on each K and T. We refer to this assumption as "parallel trends with respect to omitted variables" (PTOV). In other words, under PTOV and without any treatments ($dt = dk = 0$), the treated and control groups experience the same outcome changes $dT = dK$. PTOV requires that the heavy dashed lines in Figure 1 be parallel.

With PTOV, the relevant four first partial derivatives of the outcome mapping are represented as a two-by-two matrix S:

$$S = \begin{pmatrix} s_{Tt} & s_{Tk} \\ s_{Kt} & s_{Kk} \end{pmatrix} = \begin{pmatrix} \partial T/\partial t & \partial T/\partial k \\ \partial K/\partial t & \partial K/\partial k \end{pmatrix} \quad (1)$$

The first entry in S is the effect s_{Tt} of a unit treatment t on the treated, which is commonly known as ToT as labeled as such in Figure 1. The final entry s_{Kk} is the analog of ToT for the controls. The off-diagonal elements reflect spillovers, sometimes known as indirect effects of treatments. The spillover effect shown in Figure 1 is s_{Kt}.

S's first column difference and first row sum are central to our interpretation of difference-in-differences. We therefore establish the following definitions:

$$DiD \equiv s_{Tt} - s_{Kt} \quad (2)$$

$$\varepsilon \equiv s_{Tt} + s_{Tk} \quad (3)$$

Definition (2) is our representation of difference-in-differences (more literally, a difference in treatment derivatives) under the aforementioned parallel-trends assumption. *DiD* subtracts the effect, measured per unit *t*, of the treatment *t* on controls from its effect on the treated. The definition (3) refers to the "scale effect," which is the effect on the treated group of applying the treatment uniformly across the entire population, or what we call "the entire market." The scale effect ε is often the parameter of interest.

The case of no spillovers has *S* as a diagonal matrix ($s_{Tk} = 0 = s_{Kt}$), with no difference between *DiD* and ε or *DiD* and ToT While not ruling out the zero-spillover case, the purpose of this paper is to link the off-diagonal elements to the diagonal elements and to results from price theory. More generally, the difference between the scale effect and *DiD* is the sum of the spillover elements of *S*: $\varepsilon - DiD = s_{Tk} + s_{Kt}$.[4]

Another restriction on the *S* matrix also resembles parallel trends and drives many of our results. Specifically, administering the treatment uniformly to both the treated group and control group should not affect the difference between their outcomes:

$$s_{Tt} + s_{Tk} = s_{Kt} + s_{Kk} \tag{4}$$

We refer to assumption (4) as "Parallel Trends for Parallel Treatments" (PTPT). At this point, PTOV is distinct from PTPT. Proposition 1 establishes that the familiar procedure of dividing differential outcome changes by a treatment differential yields *DiD* if and only if the PTPT assumption (4) holds.

PROPOSITION 1 (Differential and parallel treatments). Assume parallel trends for omitted variables (PTOV) and that *dk* is neither 0 nor equal to *dt*. Then the PTPT assumption (4) is equivalent to (5):

$$DiD = \frac{dT - dK}{dt - dk} \tag{5}$$

Proof. To obtain an expression for the numerator in (5), totally differentiate $T(t,k;\theta_T) - K(t,k;\theta_K)$. With PTOV eliminating the θ terms, the numerator is *DiD* (*dt*–*dk*) plus the product of *dk* and the difference between the LHS and RHS of (4). With $dk \neq 0$, the RHS of (5) differs from *DiD* if and only if equation (4) is satisfied. QED

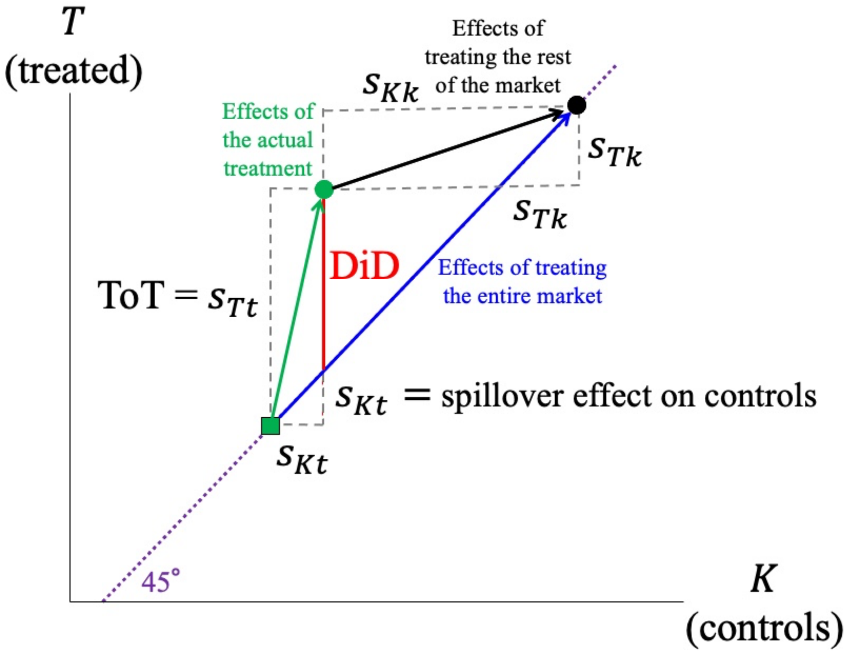
A corollary to Proposition 1 is that, with PTPT, equation (5) corresponds to the *DiD* defined in (2) regardless of whether treatments are solely for the treatment group ($dt \neq 0 = dk$) or solely for the control group ($dt = 0 \neq dk$).

[4] See also Munro et al.'s (2021) expression of a "global treatment effect" as the sum of "direct" and "indirect" treatment effects.

Figure 2 illustrates the model (1)-(4) for the case that $\varepsilon > DiD$, showing all four elements of S.[5] The axes measure outcomes for controls and treatments. The square is the baseline, showing outcomes absent any treatment. The green vector is the first column of S, showing the effects on both groups of treating only the treated. As shown, that vector is not vertical but has a slope greater than 45 degrees, which indicates that t has a spillover effect, although one that is less than the direct effect on the treated group. Unsurprisingly, the *DiD* (red segment) measures the distance between the treatment effect and the 45-degree line.

Figure 2. Geometry of Market Spillovers

in the outcomes dimension, with the scale effect exceeding *DiD*



The black vector, which is the second column of S, shows the effect of subsequently treating the rest of the market. The sum of the two arrows follows the 45-degree line if and only if the PTPT assumption (4) holds. The vertical and horizontal dimension of their sum is ε.

Figure 3 illustrates a case with the same scale effect as Figure 2, but with *DiD* closer to zero. In contrast, Figure 4's case has the same DiD as Figure 2 but no scale effect (treating the controls "undoes" the effects of t).[6] As such, it also shows an instance of $\varepsilon < DiD$.

[5] It also shows $DiD > 0$, although for what follows the sign of each ε and *DiD* is less important than the sign of their difference.

[6] The area of the triangle shown in Figures 2-4 is half of the magnitude of $DiD*\varepsilon$.

Figure 3. Market does not easily deviate from 45°

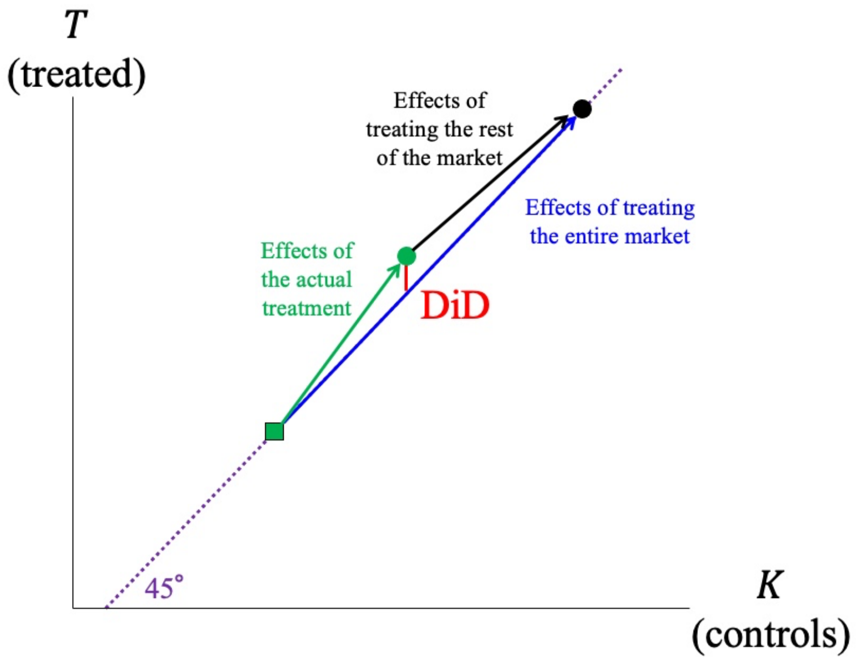
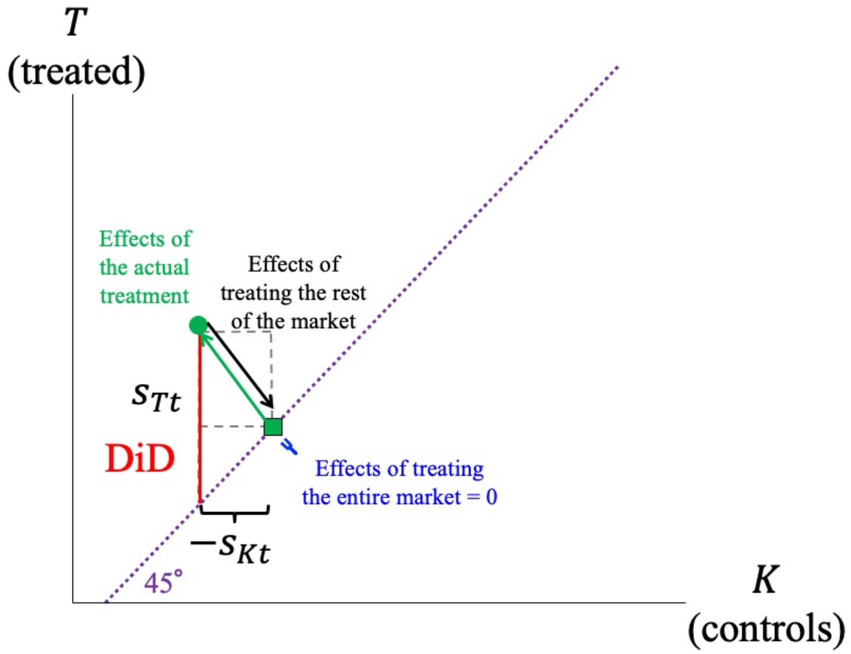


Figure 4. Scale effect = 0 << DiD



Let λ denote the share of the between-group spillover effects of a full market treatment $dt = dk$ that would be experienced by the control group.

$$\lambda \equiv \frac{s_{Kt}}{s_{Kt}+s_{Tk}} = \frac{s_{Kt}}{\varepsilon - DiD} \tag{6}$$

The symmetry of our discussion of treated and controls suggests that λ would be closely related to τ, if not identical to it, because a larger treatment group is expected to have a greater effect on the controls than treating a comparatively group would. However, until we say more about the units of *K*, *T*, *k*, and *t* (see subsection III.B and following), a precise relationship between λ and τ cannot be specified. Regardless, the common intuition that small-scale treatments ($\tau \approx 0$) have near-zero spillover effects on the controls can be represented by assuming ($\lambda \approx 0$).

PROPOSITION 2 (Treatment effects decomposition). If the PTPT assumption (4) holds, then the treatment effects matrix *S* can be written in terms of *DiD*, ε, and λ, as defined in (2), (3), and (6):

$$S = \begin{pmatrix} \lambda\varepsilon + (1-\lambda)DiD & (1-\lambda)(\varepsilon - DiD) \\ \lambda(\varepsilon - DiD) & (1-\lambda)\varepsilon + \lambda DiD \end{pmatrix} \tag{7}$$

Proof. The share defined by (6) distributes the sum of the spillover terms, already established to be $\varepsilon - DiD$, between its two components as shown in (7). From (2), the s_{Tt} term is the sum $DiD + s_{Kt}$ and therefore what is shown in (7). PTPT requires that $s_{Kk} = DiD + s_{Tk}$, which is the result shown in (7). QED

The definitions and axiom (2)-(6) allow the diagonal of *S* to be expressed entirely in terms of weighted averages of *DiD* and ε, using λ and $1-\lambda$ as weights. The off-diagonal "spillover" elements are the difference between the scale effect and DiD, scaled by either λ or $1-\lambda$. The direction of the spillover effects can therefore be understood as a comparison between the scale effect and *DiD*. The expression (7) and the intuition about signing market spillovers are familiar from price theory, where they are known as Marshall's Laws of Derived Demand.

The eigenvalues of *S* are simple, of intrinsic interest, and useful for establishing additional results.

COROLLARY. Under the PTPT assumption (4), the eigenvalues of *S* are *DiD* and ε. The matrix sum (product) of two matrices each of the form (7) itself has the form (7), with one eigenvalue that is the sum (product) of the two component *DiD*s and another eigenvalue that is the sum (product) of the two εs, respectively.

Proof. Use (7), which requires (4), to calculate the eigenvalues. QED

Even though (so far) the matrix *S* has three degrees of freedom, its eigenvalues are independent of the spillover share λ. Each eigenvalue is of intrinsic interest because *DiD* is commonly measured while ε represents a meaningful counterfactual.

III.B. Derived-demand interpretations of treatment and treatment effects

One interpretation of t and k is as log prices on the demand side of the market. T and K represent the per-capita log quantities demanded of the two corresponding goods, out of a total of N consumption choices. Final consumers have convex preferences represented by $u(x_1, \dots, x_{N-2}, e^T, e^K)$ and face a linear budget constraint. In this context, the treatments t and k result from shocks to the supply of T and K. We refer to this interpretation as demand-price treatments with quantity outcomes.

Of course, the Marshallian demand functions for the T and K goods depend on income and all N prices. In our notation, income and the N–2 prices are part of the "other factors" θ_T and θ_K. The parallel trends for omitted variables (PTOV) assumption therefore requires that the demand for the T and K goods have the same income elasticity and the same cross-price elasticities with respect to the other N–2 prices. In this context, the second parallel trends assumption (PTPT or (4)) requires that the relative quantities demanded of the T and K goods is invariant to equi-proportional changes in the two corresponding prices. Proposition 3 establishes that, due to demand-theory restrictions, the PTOV assumption guarantees PTPT, which is required for many of our previous results.

PROPOSITION 3. If the N-good demand system satisfies PTOV, then it satisfies PTPT.

Proof. The sum of the N Marshallian price elasticities for the T good must be the same as it is for the K good because the two have the same income elasticity and Marshallian demands are homogeneous. PTOV requires that the N–2 outside good price elasticities be the same for the T and K goods. With the elements of the S matrix interpreted as Marshallian price elasticites, the PTPT assumption must be satisfied because the LHS (RHS) of equation (4) is the sum of the remaining two prices elasticities from the T-good (K-good) demand function, respectively. QED

In the language of production theory, PTOV requires that the preference function exhibit homogeneous weak separability in the T and K goods.[7] That is, the two are inputs into a composite index F that is a homogeneous function of the input quantities, as shown in (8). This is why Marshall's Laws of Derived Demand are relevant.

$$\tilde{u}\big(x_1, \dots, x_{N-2}, F(e^T, e^K)\big) = u(x_1, \dots, x_{N-2}, e^T, e^K) \tag{8}$$

In some applications, as with the farmers and barbers discussed at the beginning of this paper, T and K are aptly described as production factors as Marshall (1895) does. They could be labor in two different states, as in the minimum wage literature. They could be capital in two different industries. In other applications, T and K 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, Minton, Mulligan, & Murphy (2019, Chapter 17). The model (8) is flexible in accommodating these cases.

[7] PTOV allows F to be homothetic but not homogeneous of degree one in its inputs, in which case ε should be understood as the price elasticity of F demand ε^d times a returns-to-scale factor. See Solow (1955, p. 104).

Under this price-treatment quantity-outcome interpretation, the matrix S shown in (1) becomes the two-by-two submatrix of u's Marshallian cross-price elasticity matrix S^d corresponding to the T and K rows and columns, expressed in elasticity format. The PTPT assumption (4) is automatically satisfied. Most important, the scale effect ε, the spillover share λ, and the *DiD* have precise economic interpretations, as established by Proposition 4.

PROPOSITION 4 (Hicks-Marshall). If the treatment effects matrix S is the submatrix S^d of Marshallian cross-price elasticities, and the demand system satisfies PTOV, then

(i) The scale effect ε is the Marshallian own-price elasticity of the demand $\varepsilon^d < 0$ for the composite F,

(ii) λ is the share of the combined expenditure on the T and K goods that is spent on T,

(iii) –*DiD* > 0 is the elasticity σ^d of input substitution the composite index F from equation (8), and

(iv) –*DiD* is also the shadow elasticity of substitution between the T and K goods in $u()$ as defined by McFadden (1963).

Proof. Marshall (1895) and Hicks (1936) famously prove that price elasticities of input demands satisfy (7) with the economic interpretations cited in items (i)-(iii) of the proposition. See also our Appendix I. QED

Simply put, the scale effect ε is the price elasticity of the demand for the composite because the index function F is homogeneous. Because the off-diagonal elements of S^d are cross-price elasticities, Hicksian symmetry and equal income elasticities require that the spillover share λ equal the share of the combined expenditure on the T and K goods that are spent on T. This conforms with the usual intuition that a treatment t has little effect on the control outcome K if the treatment share τ were close to zero.

Perhaps the most important result is that –*DiD* > 0 is the elasticity σ^d of factor substitution in F. As such it has no obvious relation to the scale effect and other properties of $\tilde{u}$. A treatment affects the relative price of the T and K goods, changing their ratio – a difference in logs – according to the elasticity of substitution. None of our analytical results require that elasticities or shares are constant even though our prose may refer to them as "parameters."[8]

Consumer theory permits either sign for $\varepsilon^d + \sigma^d$ and therefore either sign for the spillover elements in (7). In a complements case, as in Figures 2 and 3, the scale effect exceeds the substitution effect. That is, the magnitude of the scale effect is underestimated by *DiD*'s magnitude. At the other extreme, Figure 4 shows a case of treatment-control substitution in which the scale effect is zero.

Complementarity is the case when the treated are affected more by a full-market treatment than receiving the same treatment while others in the market are untreated. Note that complementarity requires neither increasing returns nor externalities. It does not require that the treated and controls ever meet each other to trade. It does not require Leontief preferences or even

[8] The shares and elasticities represent the values applicable to the point $\{x_1, \dots, x_{N-2}, T, K\}$ where expression (8) is evaluated.

$\sigma^d < 1$. Complementarity could even occur through an income effect on the demand for the *T* and *K* goods because ε^d includes the income effect of price treatments.

Other than ruling out $dt = dk$, this analysis of quantity outcomes does not restrict the supply of *T* and *K*. We interpret *dt* and *dk* as results of treatments, leaving implicit the supply-demand equilibrium determination of the quantitative relationship between them. A subsidy paid to the treated, for example, may move the market further up a *T*-supply curve than it moves down the *T*-demand curve represented by S^d. We leave that part implicit because the econometrics of supply-demand feedback is already well studied. The equations that we do show are valid regardless of the details of that feedback, and are adequate to show the precise relationship between scale and substitution effects on the demand side.[9]

Other DiD studies feature quantity 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.[10] These can be analyzed by inverting the Marshallian cross-price elasticity matrix S^d that maps price treatments into quantity outcomes. From equation (7), the diagonal elements of $(S^d)^{-1}$ can be expressed as a weighted average of $1/\varepsilon^d$ and $-1/\sigma^d$. Its off-diagonal spillover elements rescale the difference between these two eigenvalues.[11] The weights and scaling factors are the same two expenditure shares that apply to the case of quantity outcomes and price treatments.

By definition, DiD exaggerates the magnitude of the scale effect if and only if $\varepsilon > DiD > 0$ or $\varepsilon < DiD < 0$. If that "substitutes" case describes the quantity effects of price treatments, then the price effects of quantity treatments in the same market must fall into the "complements" category. That is, the scale effect for price outcomes would exceed *DiD* in magnitude. The proof is that the price-outcome *DiD* is the reciprocal of the quantity-outcome *DiD* while the price-outcome scale effect is the reciprocal of the quantity-outcome scale effect.

III.C. Difference-in-Differences may indicate the wrong sign

Although the scale and substitution effects on quantities of price treatments are expected to have the same sign, their signs are not necessarily aligned for alternative treatments and outcomes. Take the case of productivity treatments. Let *A* measure productivity enhancements that augment the *K* factor and *B* those augmenting the *T* factor. Formally, the prices are the same but preferences are $u(x_1, \dots, x_{N-2}, e^{T+B}, e^{K+A})$. Equivalently, consumers choose e^{T+B} and e^{K+A}

[9] Scale and substitution effects can also be investigated on the supply side. In the case of a subsidy to *T*, the factor-supply price of *T*, t^s, would exceed the demand price *t* featured in this paper. If parallel trends are also satisfied on the supply side and the elasticity of substitution in supply is denoted $\sigma^s > 0$, then a DiD constructed from supply prices would satisfy $dT - dK = \sigma^s(dt^s - dk)$, as compared to a DiD constructed from demand prices.

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

[11] S^d is invertible as long as ε^d and σ^d are not zero. Recall that the eigenvalues of the inverse of a matrix are the reciprocals of the eigenvalues of the original matrix.

subject to the augmented prices e^{t-B} and e^{k-A}. The price- and productivity-treatment effects satisfy:

$$\begin{pmatrix} d(T+B) \\ d(K+A) \end{pmatrix} = S^d \begin{pmatrix} dt \\ dk \end{pmatrix} - S^d \begin{pmatrix} dB \\ dA \end{pmatrix} \tag{9}$$

A and *B* appear on the LHS of (9) because *T*+*B* and *K*+*A* enter the preference function rather than *T* and *K* alone. As the effect of price on quantity at given productivity levels, the first term on the RHS of (9) was worked out in the previous section. The final term reflects the fact that *A* and *B* reduce the effective price of the productivity-augmented *K* and *T* goods, respectively. Solving for the treatment and control outcomes and denoting the identity matrix as *I*,

$$\begin{pmatrix} dT \\ dK \end{pmatrix} = S^d \begin{pmatrix} dt \\ dk \end{pmatrix} - (I + S^d) \begin{pmatrix} dB \\ dA \end{pmatrix} \tag{10}$$

PROPOSITION 5 (Opposite signs for scale and *DiD*). If the treatment effects matrix *S* is the matrix $-(I+S^d)$ of quantity effects of productivity treatments that hold *t* and *k* constant, then the scale effect can have the opposite sign as *DiD*.

Proof. Whereas the eigenvalues of S^d are ε^d and $-\sigma^d$ and both negative, the eigenvalues of $-(I+S^d)$ are $-\varepsilon^d-1$ and σ^d-1. With either $0 < -\varepsilon^d < 1 < \sigma^d$ or $-\varepsilon^d > 1 > \sigma^d > 0$ satisfied, the scale effect and DiD, respectively, of productivity treatments have opposite signs.

The direction of the factor-demand effects of factor-neutral productivity growth $dA=dB > 0$ depends on whether consumers' demand for the composite *F* is price elastic enough to absorb the additional production that occurs without any change in factor usage. That is a comparison of ε^d to -1. In contrast, whether factor-specific productivity growth $dB > 0 = dA$ increases or decreases *T*–*K* depends on whether factor substitution in *F* is elastic or not. That is a comparison of σ^d and 1. Simply put, the scale and substitution effects of productivity treatments can have opposite signs because $-\varepsilon^d$ can be on the opposite side of one as σ^d is.

Let *p* denote the log of the average factor cost of *F*. Because *F* is homogeneous and the factor quantities minimize cost, *p* changes are related to productivity and factor-price changes by (11):

$$dp = \lambda(dt - dB) + (1 - \lambda)(dk - dA) \tag{11}$$

where, as before, λ is *T*'s share of expenditure on *F*. Now we can return to the farmers from the beginning of this paper. The hypothetical of interest is the real-wage effects *dt*–*dp* and *dk*–*dp* of productivity growth in all occupations, represented as $dA = dB > 0$. Equation (11) requires that neutral productivity growth increase real wages by the same proportion, although the allocation of the wage increase between *t* and *k* can be uneven.[12] Regardless, the average real-wage effect is independent of the elasticity σ^d of input substitution, which is absent from (11).

[12] Depending on the supply conditions for *T* and *K*, which are unrestricted by our demand model (10), relative price changes may be required to motivate proportional increases in these two quantities.

In contrast, the $dB > 0 = dA$ special case of (10) relates factor-specific productivity growth to relative factor quantities and relative factor prices. For example, in the price theory oral tradition, $dt - dk = 0$ because the two inputs are perfect substitutes on the supply side. Factor-specific productivity growth would affect relative factor quantities to the extent that $\sigma^d \neq 1$, but not relative factor prices. In other words, market-wide productivity growth increases wages generally even while biased productivity growth has little effect on relative wages among competing occupations.[13] See also Section V.

The effects on *expenditures* of price treatments are similar to the quantity effects of productivity treatments (10). Holding productivity constant, expenditure effects are described by:

$$\begin{pmatrix} d(T+t) \\ d(K+k) \end{pmatrix} = (I + S^d) \begin{pmatrix} dt \\ dk \end{pmatrix} \tag{12}$$

where T+t is the log of expenditure on the T good, and likewise for K+k. This is another case in which the *DiD* need not have the same sign as the scale effect because the former is $1-\sigma^d$ while the latter is ε^d+1.

DiD has the opposite sign of the scale effect when a treatment has a greater effect on the outcome for the untreated than for the treated. Although this configuration of the S matrix is unexpected for the effects of prices on quantities, it can easily occur as the result of productivity treatments (among others). The market may adjust to the augmentation of T's productivity by increasing K and reducing T.

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

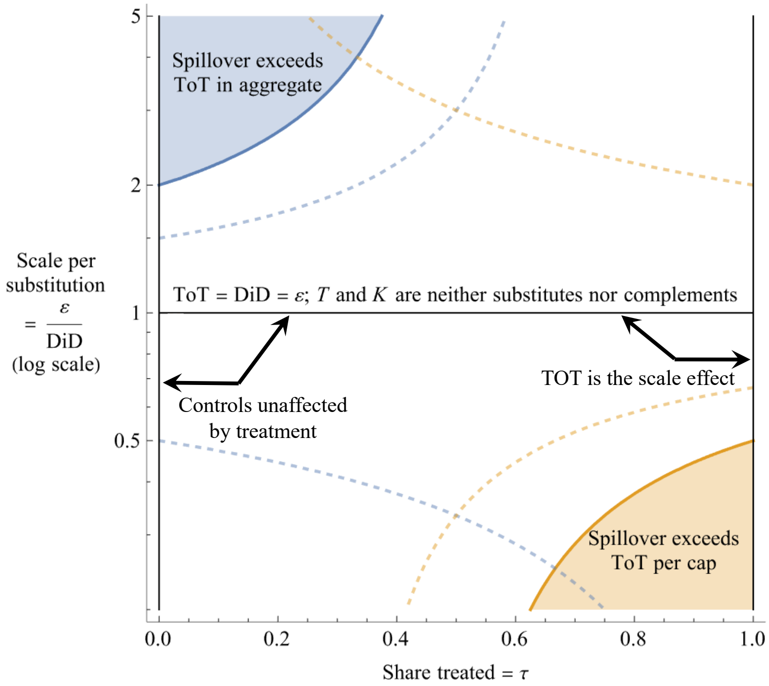
Recall from equation (7) that the effect s_{Tt} of the treatment on the treated is potentially different from *DiD*. An advantage of small treatment shares is that s_{Tt} is well approximated by *DiD*. A disadvantage is that s_{Tt} reveals little about the scale effect. It may be tempting to "scale up the treatment" and estimate the *DiD* again. But λ's relationship with *DiD* is very different than its relationship with s_{Tt}. Indeed, under the economic interpretations that tie *DiD* to an elasticity of substitution σ^d, the large-scale treatment may have the same DiD as a small-scale treatment even though the controls are far more "contaminated" in the former case.

Here we show how the share τ 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 per capita terms*. However, surprisingly, the spillover effect is comparatively large in the aggregate.

[13] Recall that *DiD* for quantity outcomes exaggerates the magnitude of the scale effect when the two inputs are substitutes. From Proposition 2's Corollary, this is exactly the case when *DiD* for price outcomes understates the magnitude of the scale effect on prices.

When the spillover share λ coincides with the treatment share τ and $dk = dA = dB = 0$, Figure 5 shows the various possibilities for the spillover effect s_{Kt} as compared to the effect of the treatment on the treated s_{Tt}. Based entirely on equation (7), Figure 5 is consistent with, but does not require any of, the consumer-theory interpretations of the treatment effects matrix S. The figure's horizontal axis is treatment share τ. The vertical axis is the ratio of the magnitude of the scale effect ε to the magnitude of DiD. The horizontal line shows all the possibilities satisfying $\varepsilon = DiD$, which (in price theory terminology) means that treatments and controls are neither substitutes nor complements.

Figure 5. The share treated and the equilibrium spillover effect



Notes: TOT = Treatment on the Treated, s_{Tt}. The dashed curves show parameter values where the spillover effect magnitude is exactly half ToT, either in aggregate (dark, $2(1-\tau)s_{Kt} = \tau s_{Tt}$) or per capita (light, $-2s_{Kt} = s_{Tt}$)

On the horizontal line, the off-diagonal spillover terms in equation (7) are zero. The other way that the spillover effect s_{Kt} can be zero is on the vertical line at zero treatment share. However, the other spillover term s_{Tk} is not zero on that line except where it intersects the horizontal line. When treatments and controls 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. The spillover term s_{Tk} is part of the effect of treating the entire market.

In price-theoretic terms, having a treatment share close to zero helps solve the control-contamination problem but at the expense of increasingly weighting the ToT s_{Tt} toward the substitution effect σ^d rather than the scale effect ε^d. ToT closely approximates the scale effect in only two circumstances: either the share treated is close to one, or the scale and substitution effects essentially cancel as they do on the horizontal line.[14]

Figure 5 also shows three light-colored curves: one solid and two dashed. The solid curve $s_{Kt} + s_{Tt} = 0$ represents those parameter combinations where the spillover effect, in per capita terms, has the same magnitude as the ToT.[15] The area below that curve represents parameter combinations where the spillover effect has, in per capita terms, the greater magnitude. These parameters are on the bottom right of Figure 5 because the spillover effect is relatively large when a large share of the market is treated, especially when the scale effect is large.

The effect of the treatment share τ can be illustrated with a geographic example. Let T represent Canada and K 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 τ near zero, a public policy implemented in Canada alone ($dt > 0 = dk$) 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 *dt* effect but not the *dk* effect that Canada would experience if the ROW were treated.

In per capita 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.[16] 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 5's solid dark curve represents those parameter combinations where the spillover effect has the same magnitude as the ToT in aggregate. Surprisingly, a reduced treatment share results in a greater proportionate decline in the magnitude of the aggregate ToT than in the aggregate spillover effect, if the latter is affected at all. The more significant aggregate spillover effects are therefore shown in Figure 5 in the lower-left corner and, especially, the upper-left corner.

The reason for this surprising result can be understood by considering the scale and substitution effects separately. A scale effect by itself (in the consumer choice context, an income effect is an example) has aggregate effects in proportion to the shares $1-\tau$ and τ. That is merely the aggregate counterpart to the parallel trends assumption that scale contributes equal *per capita* changes to both treatments and controls. With a small treatment share, essentially all the aggregate effect of scale is on the untreated. In this case, the only way to conclude that aggregate spillovers

[14] *DiD* itself may vary with τ, but that does not necessarily restrict the ratio DiD/ε.

[15] If the variable *T* and *K* represent natural logs of quantities, then s_{Kt} and s_{Tt} are interpreted as percentages.

[16] The sign comparison of the TOT and spillover effect depends on whether the varieties are substitutes (opposite sign; bottom half of Figure 5) or complements (same sign; top half of Figure 5).

are comparatively small is for the scale and substitution effects on the untreated approximately offset, as they do near Figure 5's horizontal line representing "neither substitutes nor complements."

Without any scale effect ($\varepsilon = 0$), the aggregate spillover effect would be equal to the aggregate ToT effect, albeit in the opposite direction, regardless of τ. Figure 4 shows that case. In the consumer demand context, equal and opposite aggregate effects essentially define a substitution effect.

Algebraically, the necessary and sufficient condition describing the upper-left area in Figure 5 is:

$$\left[\frac{1-\tau}{\tau}\frac{\tau(\varepsilon - DiD)}{\tau\varepsilon + (1-\tau)DiD}\right]^2 > 1 \Leftrightarrow \left(\frac{1}{2}-\tau\right)\frac{\varepsilon}{DiD} > 1-\tau \tag{13}$$

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 (13) refers 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 ($\tau < ½$) and T and K to be complements ($\varepsilon/DiD > 1$). As the share treated becomes small, the inequality (13) reduces to $\varepsilon/DiD > 2$.

Suppose, for example, that $\varepsilon = -2$ and $DiD = -1/2$. In aggregate, the direct effect of dt is $-\tau(3\tau+1)/2$ and the spillover effect is $-\tau(1-\tau)3/2$. The spillover effect has greater magnitude for any $\tau < 1/3$. In the limit as τ approaches zero, the spillover effect is three times the direct effect. Section V further illustrates this possibility in the case of a randomized labor-market experiment.

IV. Bias correction

The straightforward case for generalizing a DiD estimate is when we are interested in ToT – the effect on T of $dt > 0 = dk$ – rather than the effect of a hypothetical aggregate treatment. Proposition 2 shows that ToT = $\lambda\varepsilon + (1-\lambda)DiD$, which we expect to be similar to the DiD estimate when the share treated is close to zero. For example, one 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 $DiD\ dt$.

Even when the parameter of interest is involves the effect ε 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

Due to the invisible hand, the controls may be "contaminated" by T's treatment. Nevertheless, our results show how a DiD estimator can be useful in recovering the scale effect ε. To see this consider the total derivative of $T(t,k;\theta_T)$:

$$dT = \frac{\partial T}{\partial \theta} d\theta_T + \varepsilon \frac{dt + dk}{2} + \left[\left(\frac{1}{2} - \lambda \right) \varepsilon - (1 - \lambda) DiD \right] (dk - dt) \tag{14}$$

In words, equation (14) separates the price effects on *T* into two terms: (i) an average treatment term whose coefficient is the scale effect *ε*, and (ii) a correction term accounting for inequality of treatments. Calculating the correction term is facilitated by having an estimate of *DiD*. Although *DiD* does not include the scale effect, it can be a tool for estimating the scale effect by providing quantitative information about the amount of substitution between treatments and controls.

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

Here we briefly consider the case in which a fraction *n* of the controls is beyond the reach of the invisible hand. Outside-market controls would be contaminated neither by *t* nor *k*. The treatment effects matrix *S* becomes a weighted average of (1) and the version of it without spillovers:

$$S = (1 - n) \begin{pmatrix} s_{Tt} & s_{Tk} \\ s_{Kt} & s_{Kk} \end{pmatrix} + n \begin{pmatrix} s_{Tt} & 0 \\ 0 & s_{Kk} \end{pmatrix} \tag{15}$$

We assume that the first matrix satisfies PTPT.

The scale effect is defined as before, in equation (3). We denote the difference between the *Tt* and *Kt* elements of *S* as *DiD*(*λn*):

$$DiD(\lambda n) \equiv s_{Tt} - (1 - n) s_{Kt} = \lambda n \varepsilon + (1 - \lambda n) DiD(0) \tag{16}$$

where the second equality follows from the element-by-element equations (7). Unlike *DiD*(0), *DiD*(*λn*) puts some weight on the scale effect. The scale effect coefficient is less than one both because only part of the market is treated ($\lambda < 1$) and because only some of the controls are outside the market ($n < 1$). Still, *DiD*(*λn*) over- or under-estimates the magnitude of the scale effect according to whether *T* and *K* are substitutes or complements, respectively. If none of the controls were in the market ($n = 1$), *DiD*(*λn*) would be the *ToT* but still differ from the scale effect because it does not include the effect on the treated of applying treatments to the untreated in their market.

Having at least some of the controls out of the market raises the possibility of recovering the scale effect from a meta-analysis. Specifically, assume that two DiDs are available from distinct markets with the same *ε* and *DiD*(0) but different shares for the out-of-market controls (*n*) or different treatment shares as reflected in *λ*. Letting subscripts denote markets, the common scale effect can be written in the two-market case as a weighted average of *DiD* from each market:

$$\varepsilon = \delta DiD(\lambda_1 n_1) + (1 - \delta) DiD(\lambda_2 n_2) \tag{17}$$

$$\delta \equiv \frac{1 - \lambda_2 n_2}{\lambda_1 n_1 - \lambda_2 n_2} \tag{18}$$

Note that the market-1 weight δ must be outside the unit interval and therefore put negative weight on one of the DiDs and a coefficient greater than one on the other. One, but not both, of the markets could have $n = 0$.

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

V.A. The union wage effect

The union wage effect has been studied with the DiD method, with log wages as the outcome. In our notation, the outcome for the treated is T, which is compared to the log wage rate K for the non-union "controls." Here we interpret the unionization "treatment" as a restriction t on the supply of labor in the unionized sector to raise wages T in that sector. Licensing requirements are examples (Lewis 1963).

Early work on the union wage effect indicated a "strong presumption" of equilibrium spillovers, which are less discussed in the more recent literature.[17] Here we sketch a simple version of the labor-market equilibrium described by Rees (1962) and Lewis (1963). Union and nonunion employment totals are denoted U and N, respectively. Workers unable to gain employment in the unionized sector are employed in the nonunion sector, which means that $U = \tau - t + k$, $R = 1 - \tau - k + t$, where t and k are quantity treatments. That is, the treatment t increases union wages by shifting employment from union to nonunion while the treatment k does the opposite.

The demand side of the market has the same structure as in our Section III.B, but with a labor market interpretation. Wage-taking employers minimizing factor costs have conditional labor-demand functions that depend on aggregate output and all factor prices. As before, σ^d denotes the shadow elasticity of substitution in demand, but now the two "goods" are union and nonunion labor. Also recall that our specification of the parallel trends assumption requires that the relative treatment and control quantities vary only with their relative factor prices, and vice versa. The corresponding two-by-two matrix of cross-price elasticities is still denoted S^d. Hicksian symmetry and parallel trends together imply that the diagonal elements of $(S^d)^{-1}$ can be expressed as a weighted average of $1/\varepsilon^d$ and $-1/\sigma^d$, where the weights are the union and non-union shares of labor income.[18]

[17] Quoted from Lewis (1983, p. 3). He adds that "the relative wage of each worker depends, not only on his union status, sex, color, schooling, experience, and like variables, but also on the extent of unionism in the whole work force…." In contrast, Freeman (1984) and others refer to "the effect of unionism" and the "true impact of unionism" without drawing distinctions between DiD, TOT, etc.

[18] The demand-side scale effect $1/\varepsilon^d$ of decreasing both U and R in the same proportion is only hypothetical because the supply constraints prevent that from occurring in equilibrium. Because we refer to the treatments as t and k rather than the logs of U and R, the first column of the matrix S from the model (1) corresponding to the equilibrium union-wage model is $\left\{\frac{1}{\tau}\frac{1}{\sigma^d}, -\frac{1}{1-\tau}\frac{1}{\sigma^d}\right\}$. The second column is the negation of the first column, which is the version of (1) shown in Figure 4.

With baseline union and nonunion employment and factor shares equal to τ and $1 - \tau$, respectively, the wage effects of the quantity treatment *t* satisfy:

$$dT = -\frac{d \ln U}{\sigma^d} \tag{19}$$

$$dK = \frac{\tau}{1-\tau}\frac{d \ln U}{\sigma^d} \tag{20}$$

where $d\ln U < 0$ is the change in log union labor quantity resulting from the supply restriction.[19] Note that (19) and (20) imply that unionization increases log wages *T* in the union sector while reducing log wages *K* elsewhere with magnitudes governed by τ and σ^d.

The DiD estimator is the effect of the treatment on the union-nonunion wage gap:

$$dT - dK = -\frac{1}{1-\tau}\frac{d \ln U}{\sigma^d} \tag{21}$$

The DiD estimator (21) is different from equation (19), which is the treatment effect of unionization on the τ who remain unionized. The difference between the two is equation (20), whose final term quantifies the "contamination of" (or treatment spillover onto) the non-union controls.

The spillover term (20) would be near zero if the union share were close to zero. However, studies of union wages often include markets with sizeable union sectors.[20] In such cases, much of the union-nonunion wage gap may reflect a reduction in nonunion wages rather than an increase in union wages.[21] Even with a small union sector, the aggregate effects of supply constraints on non-union wages can exceed their effect on union wages (that is, the ToT) precisely because of the relatively large number of workers in the non-union sector.

Both equation (19)'s RHS and the DiD estimator (21) are different from (20)'s RHS times -1, which would be the effect of unionizing the remaining $1 - \tau$ of the workforce by restricting that supply by the same proportion. Particularly when the unionization rate is low, the effect on the wages of erstwhile nonunion workers of extending union status to all of them would be much

[19] Our comparative statics begin from the efficient allocation of labor between the sectors, where factor shares equal employment shares for a production function in which the two types of workers enter symmetrically. The more that union-sector labor supply is restricted, the more that union workers' factor share would exceed its employment share, adding an additional term to both (19) and (20) reflecting first-order aggregate deadweight costs of the restrictions.

[20] 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."

[21] See also Lewis (1963) and Heckman, Lochner, and Taber (1999). 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.

less than indicated by the union-nonunion wage gap (21).[22] These are further examples of how, in market settings, the DiD estimator differs from parameters that are potentially more interesting.

V.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 a $1 per pack tax would 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.

DiD pass-through studies found nearly one-for-one passthrough, but overlooked the possibility that retail prices in the control states were increased by the tax rates in the treatment states.[23] 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. If we interpret T as retail prices in the states with tax increases and K represents retail prices in the other states, that is the situation illustrated in Figures 2 and Figures 3. The blue arrow represents the retail-price effects of a nationwide tax increase. A national tax would increase prices more than a geographically-concentrated tax increase (green arrow), even in the states targeted by those taxes.

Another example is related to Jaffe, Minton, Mulligan, & Murphy (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 account for this, 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.

[22] In terms of Figure 4, the initial treatment effect (green arrow) is near vertical when τ is near zero. That is, unionization increases per-worker wages more for the τ than it reduces wages for the $1-\tau$. Unionizing the rest of the market returns the wage outcomes back to the baseline, which is hardly any per-worker change for the erstwhile non-union workers.

[23] Suppose that, 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. Our Appendix II provides a model of such price setting, expressing its results in the format (1). The eigenvalues of S prove to be the national pass-through coefficient (NPTC) and a weighted harmonic mean of one and that same NPTC. The weights depend on the trans-shipping costs, with very little weight on NPTC. That is, due to trans-shipping, difference-in-differences tend to show a one-for-one "effect" of tax on retail price, regardless of the NPTC.

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 as 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). For example, complementary businesses such as restaurants, lodging, and parking were opened nearby. But later studies found that most, if not all, of the additional employment was pulled in from other localities (Dwyer and Forsyth 2009).

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 ease (or difficulty) between villages as from outside the village economy as a whole. In our notation, that condition is $\varepsilon = DiD$.

V.C. Welfare effects of random treatments

Let's examine treatment effects in a straightforward substitutes setting. As such, the substitution effect exceeds the scale effect. A large pool of *ex ante* identical workers supplies hours on the intensive margin. Their population is normalized to one. From employers' perspective, any worker's hours are perfect substitutes for another's. In the baseline, each worker is paid the same hourly wage w and supplies the same hours. The aggregate demand for their hours is $D(w)$, with $D'(w) < 0$. The per capita supply of labor is $L(w)$, with $L'(w) > 0$.

An experiment selects a fraction τ of the workers for a wage subsidy $t \geq 0$.[24] Their hours are denoted T per treated and τT in total. The untreated "controls" supply K per control and $(1-\tau)K$ in aggregate. To highlight the analogy with the model (1), our notation also includes k as a subsidy for the controls, although it is not emphasized here. Given values for τ, t and k, an equilibrium is a list $\{w,T,K\}$ of wage and hours satisfying:

$$K = L(w + k)$$

$$T = L(w + t)$$

$$(1 - \tau)K + \tau T = D(w)$$

[24] This is a simplified version of Heckman, LaLonde, and Smith (1999) that focuses on incidence rather than employment effects.

Unsurprisingly, $dK/dt < 0 < dT/dt$ and $dw/dt < 0$.[25] The subsidy benefits the treated and employers (even those who do not employ any treated) and harms the controls.[26] Regardless of the share treated, the first magnitude can easily be less than the combined magnitudes of other two.

Take the case when labor demand is wage inelastic, $\tau \in (0,1)$, and the subsidy is small.[27] The treated benefit from the subsidy, but employers benefit even more because they pay less for both treated workers and untreated workers. Shrinking the treated share does not change this result. If the treated are to be the primary beneficiary of the subsidy, demand needs to be wage elastic enough or supply inelastic enough.[28] Clearly, quantifying the scale effect is essential for understanding the relationship between the welfare effect on the treated and welfare effects more broadly.

This example also distinguishes the DiD estimator $dT/dt - dK/dt$ from the effect of treating all workers. The DiD estimate is $L'(w)$ because the subsidy moves treated and controls in opposite directions along the supply curve. The equilibrium quantity effect of subsidizing all workers, is a parameter of interest and results from shifting the supply curve downward by dt. As expected from the general substitutes case, this scale effect is closer to zero than the DiD estimate.

VI. 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. In our framework, parallel trends require the treatment and control outcomes to be weakly separable in utility, production, or cost from all other outcomes. Marshall's Laws of Derived Demand are thereby vehicles for several analytical results. One is that a DiD estimator measures the degree of substitution between treatments and controls, regardless of the fraction of the market that is treated and the magnitude of market spillovers (Proposition 4). In contrast, the effect of treating the entire market is a "scale effect," which is the degree of substitution with goods outside the market where treated and controls participate. The effect of the treatment on the treated

[25] For this example, the matrix elements corresponding to (1) are $s_{Tt} = L'(w)\frac{(1-\tau)L'(w)-D'(w)}{L'(w)-D'(w)}$, $s_{Tk} = -(1-\tau)L'(w)\frac{L'(w)}{L'(w)-D'(w)}$, $s_{kT} = -\tau L'(w)\frac{L'(w)}{L'(w)-D'(w)}$, and $s_{Kk} = L'(w)\frac{\tau L'(w)-D'(w)}{L'(w)-D'(w)}$. These satisfy $\lambda = \tau$ and the parallel trends assumption (4).

[26] The expressions for aggregate effects on surplus for treated, controls, and employers are $(dw+dt)\tau T$, $(1-\tau)Kdw$, and $-D(w)dw$, respectively.

[27] A "small subsidy" refers to the comparative static $dt > 0$ in the neighborhood of $t = 0$, holding k constant at zero.

[28] For non-zero supply and demand elasticities, the aggregate benefit for the treated as a ratio to the aggregate employer benefit is $1 - \tau - D'(w)/L'(w)$.

(ToT) is a weighted average of the scale effect and the DiD, whereas the market spillovers are proportional to their difference (Proposition 2).

Proposition 2 also establishes that, assuming "parallel trends for parallel treatments" (PTPT), the eigenvalues of the treatment-effects matrix are the scale effect and the DiD. As a result, any arithmetic operations on treatment-effects matrices translate into the same operations on their respective scale effects and *DiD*s. This correspondence appeared in a few of our examples where treatment effects on demand were inverted, combined with an identity matrix, or combined with treatment effects on supply.

The presumption that market-equilibrium responses are typically dampened relative to experimental evidence (an example of which is provided in Banerjee and Duflo (2009)) may refer to the market-level feedback between supply and demand, which we have left only implicit in this paper. For the usual incidence reasons, for example, the market-level reduced form for the quantity elasticity of tax changes is $-\varepsilon^s\varepsilon^d/(\varepsilon^s-\varepsilon^d)$. This incidence coefficient reflects equilibrium dampening in the sense that it is less than both the demand elasticity magnitude $-\varepsilon^d$ and the supply elasticity ε^s.[29]

However, there is more to market equilibrium than supply-demand feedback. In particular, the actions of market participants—even those on just one side of the market—are coordinated by prices. Either the controls are affected by the treatment, the treated would be further affected by treating the rest of the market, or some combination thereof. Market spillovers drive a wedge between *DiD* and the scale effect. The market-level demand elasticity ε^d, the market-level supply elasticity ε^s and the market-level incidence coefficient $\varepsilon^s\varepsilon^d/(\varepsilon^s-\varepsilon^d)$ are each examples of a scale effect. The DiD from an experiment with control and treatment in the same market recovers substitution effects instead of scale effects. Treatment-control comparisons by themselves do not even partially identify ε^d, ε^s, or $\varepsilon^s\varepsilon^d/(\varepsilon^s-\varepsilon^d)$.

Complementarity is the case when the treated are affected more by a full-market treatment than by receiving the same treatment while others in the market are untreated. Note that complementarity requires neither increasing returns nor externalities. It does not require that the treated and controls ever meet each other to trade. It does not require Leontief preferences or technology. Complementarity in this sense only means that the scale effect exceeds the substitution effect.

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.

Per capita market 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

[29] The incidence coefficient can be derived in the usual way as the equilibrium quantity effect of a one price-unit wedge between market supply and market demand.

disadvantages. One is that scale effects are especially obscured by substitution effects. Second, and surprisingly, the spillover effect is comparatively large in the aggregate.

Appendix I: Derivation of the Hicks-Marshall Laws of Derived Demand

As with equation (8), this appendix interprets T and K as log quantities and t and k as log prices. In levels, the conditional demand equations for the T and K goods are:

$$e^T = \frac{\partial C(e^t, e^k, Y)}{\partial e^t} \wedge e^K = \frac{\partial C(e^t, e^k, Y)}{\partial e^k} \tag{22}$$

where the cost function C is the minimum expenditure of achieving output Y:

$$C(e^t, e^k, Y) \equiv \min_{T,K} e^{T+t} + e^{K+k} \quad s.t. \quad F(e^T, e^K) = Y$$

Here λ is the T good's share of C. As usual, C is homogeneous of degree one in prices and of degree one in output. The elasticity of substitution in F, which we denote $\sigma^d > 0$, is defined as the cross-price derivative of $C(w,r,Y)$ times $C(w,r,Y)/[(1-\lambda)\lambda]$. That makes the cross-price elasticity of either conditional input demand equal to the product of σ^d and the other input's share. By homogeneity, its own price elasticity is the negation of its cross-price elasticity. From (22), the log-derivative form of T and K demand are therefore:

$$dT = d\ln Y - (1-\lambda)\sigma^d(dt - dk) \wedge dK = d\ln Y + \lambda\sigma^d(dt - dk) \tag{23}$$

Let $\varepsilon^d < 0$ denote the Marshallian price elasticity of demand for F associated with the preferences $\tilde{u}$ shown in equation (8). If income and the other prices are constant, then $d\ln Y = \varepsilon^d dp$, where e^p is the average and marginal price of F. Equations (11) (without the two productivity terms) and (23) then require (24):

$$\begin{gathered} dT = [\lambda\varepsilon^d - (1-\lambda)\sigma^d]dt + (1-\lambda)(\varepsilon^d + \sigma^d)dk \wedge \\ dK = \lambda(\varepsilon^d + \sigma^d)dt + [(1-\lambda)\varepsilon^d - \lambda\sigma^d]dk \end{gathered} \tag{24}$$

which coincides with equation (7) when ε is replaced with ε^d and *DiD* with $-\sigma^d$.

Recall that Section III.B defines the matrix S^d in terms of u's Marshallian cross-price elasticity matrix. By definition (2), *DiD* is the difference between two of those elasticities, which by the Slutsky equation and equal income elasticities is also a Hicksian price elasticity difference. Because Hicksian cross-price elasticities are proportional to Allen (1938) partial elasticities of substitution (of u), *DiD* is also the difference between Allen elasticities up to the same expenditure-share proportion. With PTPT and McFadden's (1963) definition of the shadow elasticity of substitution in terms of expenditure shares and Allen substitution elasticities, –DiD must also be the shadow elasticity of substitution between the T and K goods in u.

Appendix II: National and local pass-through of excise taxes

The outcomes are retail prices *T* and *K*, expressed in levels (as in the literature). The treatments are excise tax rates *t* and *k*, respectively. Consumers are not mobile between the treatment and control areas, which have populations τ and $1-\tau$, respectively. Per-capita consumer demand in each area is a function *D*() of the area's retail price. Supply prices are *T–t* and *K–k*, respectively.

A national manufacturer sets retail prices in each area to maximize profits:

$$\begin{aligned}&\big[T - t - c\big((T-t)-(K-k)\big)\big]\tau D(T)\\&\qquad + \big[K - k - c\big((T-t)-(K-k)\big)\big](1-\tau)D(K)\end{aligned} \tag{25}$$

where the average and marginal cost *c*() depends on the gap in supply prices between areas. It is a convex function and minimized when the two areas have the same supply price. These assumptions reflect incentives for trans-shipping by area wholesalers, who reimburse the supply price to the manufacturer and handle excise tax payments. Especially, supply-price gaps incentivize wholesalers in the low-price areas to acquire more quantity than needed for their area and (before excise tax is determined) sell the excess to wholesalers in other areas.

The first order conditions for maximizing profit are:

$$\begin{aligned}&\tau\big\{D(T) + \big[T - t - c\big((T-t)-(K-k)\big)\big]D'(T)\big\}\\&\qquad = [\tau D(T) + (1-\tau)D(K)]c'\big((T-t)-(K-k)\big)\\&\qquad = (1-\tau)\big\{D(K) + \big[K - k - c\big((T-t)-(K-k)\big)\big]D'(K)\big\}\end{aligned} \tag{26}$$

By totally differentiating the two equations (26) in the neighborhood of $t = k$ while holding τ constant, yields expressions for *dT* and *dK* as functions of *dt* and *dk*. We write them in the matrix form (1) (not shown), and simplify them with two definitions:

$$\rho^d(T) \equiv \left[2 - \frac{D''(T)/D'(T)}{D'(T)/D(T)}\right]^{-1} > 0 \wedge \eta(T) = \frac{T}{D(T)}D'(T) < 0 \tag{27}$$

The first is the pass-through coefficient ρ^d defined in the usual way as a transformation of the demand function. The second is the price elasticity of demand. Both depend on the retail price because they are not necessarily constant along the demand curve.

With the *S* matrix so derived, its eigenvalues are:

$$\varepsilon = \rho^d \wedge DiD = \left[\omega + (1-\omega)\frac{1}{\rho^d}\right]^{-1} \tag{28}$$

DiD is a weighted harmonic mean of 1 and ρ^d, where the weight is $\omega \equiv \frac{Tc''(0)}{Tc''(0)-(1-\tau)\tau\eta}$. None of the weight is on ρ^d as the trans-shipping cost function becomes more convex or as the treatment share τ approaches zero. The weight λ for calculating ToT from the two eigenvalues is, in this application, the population share τ.

References

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

Allen, R. G. D. 1938. *Mathematical Analysis for Economists.* Macmillan.

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.

Banerjee, Abhijit V., and Esther Duflo. 2009. "The experimental approach to development economics." *Annu. Rev. Econ.* 1: 151–178.

Banzhaf, H. Spencer. 2021. "Difference-in-differences hedonics." *Journal of Political Economy* 129: 2385–2414.

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.

Heckman, James J., and Edward Vytlacil. 2005. "Structural equations, treatment effects, and econometric policy evaluation 1." *Econometrica* 73: 669–738.

Heckman, James J., Lance Lochner, and Christopher Taber. 1999. "Human capital formation and general equilibrium treatment effects: a study of tax and tuition policy." *Fiscal Studies* 20: 25–40.

Heckman, James J., Robert J. LaLonde, and Jeffrey A. Smith. 1999. *The economics and econometrics of active labor market programs.* Vol. 3, in *Handbook of labor economics*, 1865–2097. Elsevier.

Hicks, John. 1936. *The Theory of Wages.* London: Macmillan.

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.

Lewis, H. Gregg. 1963. *Unionism and Relative Wages in the United States: An Empirical Inquiry.* Chicago: University of Chicago Press.

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.

McFadden, Daniel. 1963. "Constant elasticity of substitution production functions." *The Review of Economic Studies* 30: 73–83.

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.

Munro, Evan, Stefan Wager, and Kuang Xu. 2021. "Treatment effects in market equilibrium." *arXiv preprint arXiv:2109.11647* (arXiv).

Rees, Albert. 1962. *The Economics of Trade Unions.* Chicago: 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.

Solow, Robert M. 1955. "The production function and the theory of capital." *The Review of Economic Studies* 23: 101–108.

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.