2010

The Welfare Costs of Market Restrictions

David Colander
Middlebury College

Sieuwerd Gaastra
Middlebury College

Casey Rothschild
Middlebury College

Citation

David Colander, Sieuwerd Gaastra, and Casey Rothschild (2010). The Welfare Costs of Market Restrictions. Southern Economic Journal 77(1), 213-223.

Southern Economic Journal 2010, 77(1), 213–223

Targeted Teaching

The Welfare Costs of Market Restrictions

David Colander,* Sieuwerd Gaastra,† and Casey Rothschild‡

In most introductory and intermediate microeconomics textbooks, the measurable welfare effects of price controls, quantitative restrictions, and market restrictions more generally, are depicted as a Harberger triangle. This depiction understates these restrictions' inefficiency costs because it captures only the ‘‘top-down’’ distortion caused by the wedge these restrictions drive between market-wide quantity demanded and quantity supplied. It ignores the ‘‘bottom-up’’ distortions caused by allocative inefficiencies on the constrained side of the market. In this article we describe a simple graphical exposition of these bottom-up distortions. We argue that this graph can provide students with a picture of both the top-down and bottom-up inefficiencies. Moreover, it can be used for simple back-of-the-envelope estimates of the magnitudes of the two inefficiencies.

JEL Classification: A2, D61, L51

1. Introduction

Many of the central ideas in economics are conveyed to students in graphs that provide a visual picture of economists' insights. One of the most well known of these pictures is the Harberger triangle, which is used to illustrate the deadweight loss from market restrictions such as monopoly power, quantity restrictions, and price ceilings and floors. Although it is generally known that the Harberger triangle misses important elements of these restrictions' inefficiency costs (see, for instance, Friedman and Stigler 1946; Glaeser and Luttmer 2003), this insight has not been integrated into economic textbooks. This is problematic because these overlooked inefficiency costs are theoretically important and in many cases are larger than the inefficiencies conveyed by the Harberger triangle.

In this short article we show why the Harberger triangle significantly understates the efficiency costs of any restriction that does not inherently direct (or provide incentives for) agents to efficiently deal with it. We then provide a simple graphical method of capturing the additional deadweight loss in the form of a second triangle that can be seen as a measure of this additional deadweight loss. This graphical method should make it easier to integrate these insights into the textbooks and thereby help remedy the deficiencies of presentations based only

* Middlebury College Department of Economics, Warner Hall, Middlebury, VT 05753, USA; E-mail colander@middlebury.edu; corresponding author.

† Middlebury College, Box 2877, Middlebury, VT 05753, USA; E-mail sgaastra@middlebury.edu.

‡ Middlebury College Department of Economics, Warner Hall, Middlebury, VT 05753, USA; E-mail crothsch@middlebury.edu.

We are thankful to Ted Bergstrom and Laura Razzolini (the editor) for making important suggestions that improved the exposition in the article.

Received November 2009; accepted November 2009.

on the Harberger triangle.[1] We focus on the example of price controls but discuss how the analysis caries over to other restrictions such as quotas. We argue that, together, the second triangle and the Harberger triangle provide students with a much better picture of the costs of these market restrictions, a better sense of the relative magnitudes of the two types of deadweight loss, and a better segue into a discussion of the costs of market restrictions.

The problem with using the Harberger triangle—the area between the demand and supply curve and between the pre- and post-control quantities—as a measure of the inefficiency resulting from, for example, a price floor is that it captures only one type of restriction-induced inefficiency: the inefficiency that arises because the market restriction prevents some mutually beneficial trades from taking place. Such inefficiencies might be called "top-down" inefficiencies because they would exist even if each side of the market consisted of a single representative agent reacting optimally to restrictions and dealing with the restriction in as efficient a manner as possible. When there are many agents affected by a price control, representative agent assumptions are inappropriate, and such controls will impose additional "bottom-up" costs on society. This bottom-up inefficiency occurs because, in addition to *preventing mutual beneficial trades*, price controls remove incentives for the *right* trades to take place. They therefore impose *wrong-trade*, bottom-up, social costs: They lead some of the wrong agents to do the supplying or demanding. A price floor, for example, both causes an inefficiently low quantity of the good to be supplied *and* fails to incentivize the lowest-cost potential suppliers to do that supplying. For example, faced with a minimum wage restriction, jobs will have to be rationed, but McDonald's and other minimum wage employers will have no incentive to ration those jobs in the most efficient manner—for example, to those who benefit the most from receiving them.

Similarly, a price ceiling can be expected to drive the highest-cost suppliers out of the market, but it fails to provide incentives to efficiently allocate the supply-limited quantity to the highest marginal benefit demanders. It (along with assorted shady political dealings) can therefore lead to Congressman Charles Rangel's (D-NY) renting and maintaining four rent-stabilized New York apartments at approximately half of their fair market value, even when other potential tenants might value those apartments significantly more highly (Kocieniewski 2008).

The problem could be partially resolved if the Harberger triangle was accompanied by a discussion of bottom-up inefficiency, but, because pictures tend to guide the discussions, that often does not happen.[2] Because bottom-up costs are not captured in the standard textbook

[1] In Appendix 1, we discuss the history of the graphical exposition presented in this article.

[2] There seems to be an inverse relationship between the use of the Harberger triangle to discuss the welfare loss and the mention of bottom-up inefficiency. This is to be expected. When one provides a graphical picture, that picture tends to guide the discussion. As such, authors face a choice between using graphs and providing a more complete discussion. For example, Hubbard and O'Brien (2007) and Frank and Bernanke (2009) use the Harberger triangle to demonstrate the welfare loss from price controls but do not discuss the bottom-up costs (Hubbard and O'Brien 2007, pp. 105–8; Frank and Bernanke 2009, pp. 180–3). Frank and Bernanke later (pp. 185–7) discuss how first-come first-serve policies are less efficient than highest reservation price allocation in airline booking, which gets at the bottom-up inefficiency, but they do not draw a parallel to allocations with price ceilings and floors. In contrast, Mankiw (1998, pp. 117–20), Baumol and Blinder (1998, pp. 83–85), and Krugman and Wells (2005, p. 92) do not use a graph indicating the Harberger triangle in their discussion of the costs of price controls but *do* at least mention inefficiencies that would fall into the category that we call bottom-up inefficiencies.

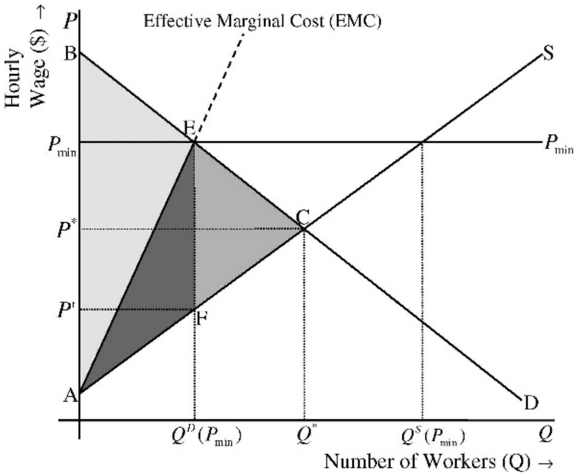

Figure 1. The deadweight loss from a price floor. The conventional top-down deadweight loss from the price floor of P_{min} is given by the Harberger triangle EFC. The additional deadweight loss due to bottom-up distortions is given by the bottom-up inefficiency triangle EFA.

graph, often the full costs of price controls are not conveyed to students. The contribution of this article is to motivate and advocate for a simple way of visually capturing these additional costs and thereby to provide a picture that will be more conducive to a broader discussion and conceptual understanding of the welfare costs of price controls and other market restrictions. We illustrate our suggested picture in the context of imposing a minimum wage in the labor market for fast food workers.

2. Illustrating the Two Types of Inefficiency: The Effective Marginal Cost Curve

Figure 1 illustrates our suggested method with an example of a market for jobs at fast food establishments such as McDonald's in the presence of a minimum wage P_{min}.

In Figure 1, the Harberger Triangle (EFC) captures the top-down inefficiency caused by the minimum wage. There are Q^* mutually beneficial trades to be made in this market, because the Q^* highest-benefit employers have higher marginal benefits than the Q^* lowest opportunity cost workers. The minimum wage prevents some of these trades from taking place by reducing the quantity demanded to $Q^D(P_{\text{min}})$. If the market could somehow ensure that this reduction occurred by removing only the highest opportunity cost workers from the market, the type of trade being eliminated would be trades such as the one between the worker just to the right of point F and the employer just to the right of point E; summing up the lost benefits from these eliminated trades would then yield the Harberger triangle.

However, because the market is not a top-down process, it typically will not lead to only the highest opportunity cost workers being rationed out of the market. This means that there will be additional costs resulting from the wrong workers receiving the jobs. Our goal is to depict these additional costs. Toward doing that, we first note that this minimum wage leads to an excess supply $[Q^S(P_{\text{min}}) - Q^D(P_{\text{min}})]$ of workers. Fast food establishments will thus get $Q^S(P_{\text{min}})/Q^D(P_{\text{min}})$ applications for each of their job postings.

If workers differ *only* in their reservation wages—so that employers are otherwise indifferent as to whom they hire—then it is reasonable to assume that jobs will be randomly allocated to willing workers. The probability that any willing worker will receive a job therefore will be $Q^D(P_{\min})/Q^S(P_{\min})$. This means that $Q^D(P_{\min})/Q^S(P_{\min})$ measures the expected *fraction* of any subset of willing workers who will actually receive jobs. In particular, because $Q^S(P)$ of the willing workers have reservation wages below any price P, $\{[Q^D(P_{\min})]/[Q^S(P_{\min})]\} \times Q^S(P)$ of *hired* workers will have reservation wages below P, at least in expectation.

This reasoning motivates the curve labeled "Effective Marginal Cost" (EMC) in Figure 1. The EMC curve is calculated by taking the supply curve and compressing it horizontally by the probability-of-receiving-a-job factor $Q^D(P_{\min})/Q^S(P_{\min})$. It therefore captures the schedule of the expected number of *hired* workers with reservation wages below any given wage P.

Inverting this reasoning allows one to think of the supply and EMC curves in terms of marginal social costs—which is why we give it the "effective marginal cost" moniker. Imagine taking the $Q^D(P_{\min})$ workers actually hired and arranging them in order of increasing reservation wages. The height of the EMC curve will then give the expected reservation wage of the Qth of these workers. In other words, the height of the EMC curve gives the expected social cost of hiring the Qth least willing-to-work *among those who were actually hired.*

This interpretation lets us use the EMC curve to compute the welfare consequences of a price floor or any similar market restriction. By construction, the EMC curve measures the expected social cost of the workers hired under the minimum wage law. The demand curve measures the social benefits of hiring workers. The area between $Q = 0$ and $Q = Q^D(P_{\min})$ — the triangle BEA in Figure 1—therefore measures the total surplus created by hiring in the presence of a minimum wage. The surplus generated without a minimum wage is measured by the area between the demand and supply curves between $Q = 0$ and the free-market equilibrium quantity—triangle ABC in Figure 1. The total deadweight loss is the difference between these two surpluses—that is, triangle AEC in Figure 1.

Note that this deadweight loss triangle decomposes nicely into the traditional Harberger triangle EFC and a second triangle AEF. We dub the latter the "bottom-up" inefficiency triangle to highlight that it results from actual bottom-up market dynamics—in particular, from the random allocation of demand-limited jobs to willing workers.[3]

To provide a concrete example, suppose that the labor supply and labor demand curves are given by $Q^S = 10P$ and $Q^D = 80 - 10P$, respectively, where P is the wage. The market-clearing wage and labor supply are then $P^* = 4$ and $Q^* = 40$. If a minimum wage of $P_{\min} = 5$ is imposed, $Q^S(P_{\min}) = 50$ workers will wish to supply labor, and only $Q^D(P_{\min}) = 30$ workers will be demanded and hired. The probability that any willing worker will get a job is thus 3/5, so the EMC curve will be given by $Q^{EMC}(P) = (3/5)\cdot(10P) = 6P$.

[3] The above example uses a linear supply curve and a linear demand curve. The reasoning and geometry generalize to nonlinear supply and demand curves in a straightforward way. The EMC curve is a "horizontally compressed" version of the supply curve, with compression factor $Q^D(P_{\min})/Q^S(P_{\min})$. The "bottom-up" inefficiency loss is measured by the area between the EMC and supply curves between $Q = 0$ and $Q^D(P_{\min})$.

The reasoning generalizes to the linear inelastic supply curves commonly studied in introductory texts in a straightforward way. If an inelastic supply curve is truly linear, it strikes the P-axis at a negative price, but the analysis is otherwise identical. If there are no suppliers willing to supply at a negative price, then the supply curve is not "truly" linear, because it has a "kink" where it flattens out at a zero price. In this case, the welfare loss would be represented by a quadrilateral with vertices at the analogs of E and F from Figure 1 and with two additional vertices at the intersections of the EMC and supply curves with the quantity axis. In other words, flattening out the supply curve at a zero price "lops off" the negative-P portion of the bottom-up inefficiency triangle that would obtain if the supply curve stayed linear below $P = 0$.

The traditional, top-down deadweight loss triangle has vertices at the unregulated equilibrium (P^*, Q^*) = (4, 40) and at the points (3, 30) and (5, 30) on the labor supply and EMC curves at the demand-restricted quantity $Q^D(P_{\min}) = 30$, respectively. The bottom-up deadweight loss triangle has vertices at the latter two points and the origin. The top-down deadweight loss is thus 1/2(5 – 3)(40 – 30) = 10, and the bottom-up deadweight loss is 1/2(5 – 3)(30 – 0) = 30.

The traditional, top-down inefficiencies occur because the imposition of the price floor eliminates some mutually beneficial trades from taking place. In particular, imposing the minimum wage removes 10 of the 40 mutually beneficial hires from taking place; when the minimum wage is in place, workers remain who would gladly work at a wage some firm would gladly pay them.

The bottom-up inefficiencies arise because, in addition to too *few* trades taking place, the *wrong* trades are also likely to take place. Here, only 30 of the 50 willing workers actually receive a job, and the recipients are unlikely to be the most efficient hires—that is, the workers with the highest net benefit of employment. Indeed, when jobs are randomly allocated to willing workers (as we assume in deriving the EMC curve), it is possible—likely, in fact—that some jobs will be allocated to workers who would not even have wanted to work at the market clearing wage $P^* = 4$.

3. Rent Seeking and Bottom-Up Inefficiency

The bottom-up inefficiency costs as we have specified them are quite separate from rent-seeking costs; our specification assumes that no rent seeking whatsoever takes place. Instead, the market dynamics implicit in our derivation of the bottom-up inefficiency are based on the assumption that the supply is rationed randomly. This assumption is likely to hold perfectly only in very particular cases, and different assumptions about rationing procedures would result in different bottom-up inefficiencies.

For example, if a frictionless secondary market facilitated retrading after the primary market had closed, the bottom-up inefficiency costs related to misallocation would be completely eliminated. Although the bottom-up efficiency costs may be eliminated in this special case, they will likely simply be replaced by rent-seeking costs as participants in the market expend socially unproductive effort in attempting to secure a favorable initial allocation in the bottom-up rationing process (Tullock 1967, p. 230).

Rent seeking and bottom-up inefficiencies are closely interconnected. For example, in the minimum wage example above, one might expect workers with lower reservation wages to expend greater rent-seeking efforts within the rationing process to increase their probability of securing a job than those with higher reservation wages. On the one hand, this would help to alleviate the misallocation cost captured in our bottom-up triangle. On the other hand, such efforts are themselves socially costly, and there should be no presumption that rent seeking will reduce *overall* inefficiencies.[4]

[4] Suppose, for example, that workers can exert equally effective—but personally costly—effort in jockeying for jobs, and that the $Q^D(P_{\min})$ highest-effort workers receive the jobs. Then we would expect the total cost of effort exerted in seeking jobs to equal the rectangle EF$P'P_{\min}$ in Figure 1, because the $Q^D(P_{\min})$ lowest reservation wage workers will each exert just enough effort to dissuade the next most willing-to-work worker from exerting any effort. (This is analogous to Posner's [1975] result about the rent-seeking costs of monopoly.) This rectangle is bigger than—in fact, exactly twice the size of—the bottom-up inefficiency triangle in Figure 1.

Turning this argument around, note that to the degree that rent seeking reduces bottom-up inefficiencies, the costs of rent seeking should be measured *net* of these bottom-up inefficiencies. This means that the standard estimates of the costs associated with rent-seeking activities (Posner 1975) may be overstated: Insofar as rent seeking reduces bottom-up inefficiencies by better allocating the production or disposition of a good, rent seeking may have some socially beneficial results.

Of course, it is not at all clear that rent seeking will always *improve* allocative inefficiencies, because the cost of improving one's standing in the rationing "lottery" will not *necessarily* be related to the value of winning it in any monotonic fashion. A connected politician would find it relatively easier to improve his standing in the rent-controlled housing "lottery" even if he had a low net benefit from winning it, for example.

In short, our point is not to argue that the Harberger triangle understates the social cost of a price floor by an amount exactly equal to the bottom-up inefficiency triangle in Figure 1. Rather, it is simply to get students to realize that other costs are occurring, and that they need to be taken into account. The bottom-up efficiency-loss triangle provides a visual segue into helping students think about the bottom-up market microstructure underlying these additional costs, and therefore it serves a central pedagogical purpose.

4. Other Examples of the Importance of Bottom-Up Inefficiencies

The price floor illustrated in Figure 1 was just one example of where the bottom-up inefficiency fundamentally changes the conceptualization of the effects of restrictions on markets. In this section we consider three other examples.

The first example is a price ceiling. The analysis of a price ceiling is entirely analogous to the preceding analysis of a price floor, except now it is the demanders who are being rationed, so there is an Effective Marginal Benefit (EMB) curve instead of an EMC curve. Assuming that each unit of total demand $Q^D(P_{\max})$ is equally likely to receive the good (for example, if there is random allocation and each individual demands at most one unit), the probability that any given unit of demand will be satisfied is $Q^S(P_{\max})/Q^D(P_{\max})$. The number of satisfied demand units whose marginal benefit from receiving the good is greater than any given price P (above $P_{\max}$) is therefore given by $\{[Q^S(P_{\max})]/[Q^D(P_{\max})]\} \times Q^D(P)$, and the EMB curve is simply a horizontally compressed version of the demand curve, with a compression factor of $Q^S(P_{\max})/Q^D(P_{\max})$. For a linear demand curve, this means that it has the same P-axis intercept, but it is more steeply sloped by the factor $Q^D(P_{\max})/Q^S(P_{\max})$. The deadweight loss from the price ceiling is then given by the triangle with vertices at the free market equilibrium, the P-intercept of the demand curve, and the intersection of the supply and EMB curves. This is depicted in Figure 2 for the special case of a perfectly inelastic supply curve.

For a concrete example, suppose the supply curve depicted in Figure 2 is vertical at $Q^S = 5$, and the demand curve is given by $Q^D(P) = 10 - P$. Then the market-clearing price is $P^* = 5$, and the imposition of a price ceiling $P_{\max} = 3$ leads to a demand of $Q^D(P_{\max}) = 7$ and a shortage of 2. If the limited quantity supplied is allocated randomly to demanders, then the probability that any unit of demand will be met is 5/7, so $Q^{EMB}(P) = (5/7)\cdot(10 - P)$ describes the EMB curve.

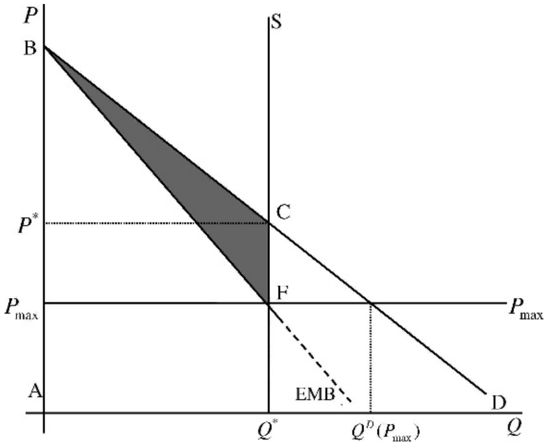

Figure 2. The deadweight loss from a price ceiling, with perfectly inelastic supply. Triangle BFC shows the total deadweight loss resulting from the bottom-up distortions induced by a price ceiling in a market with a perfectly inelastic supply curve. The inefficiency is entirely due to bottom-up distortions, because the conventional deadweight loss from price ceiling is zero.

The price ceiling induces a total deadweight loss that is measured by the bottom-up triangle BCF in Figure 2. Vertex B lies at the *P*-axis intercept of the demand (and the EMB) curve, or (0,10) in this example. Vertices C and F lie at the points (5, 5) and (5, 3), respectively, the points on the demand and EMB curves at the supply-restricted quantity $Q^S(P_{\max}) = 5$. The price ceiling thus induces a deadweight loss $1/2(5 - 3)\cdot(5)=5$.

We chose this special case because it illustrates how misleading the conventional graphical analysis of price controls can be. The conventional analysis of the imposition of a price ceiling suggests to students that when there is a perfectly inelastic supply there is no efficiency loss from a price ceiling, because the quantity supplied does not change and the Harberger triangle is nonexistent. That would be correct if only top-down inefficiency were considered, but it is not correct when there is bottom-up inefficiency. As Figure 2 and our numerical example illustrate, the price ceiling creates excess demand equal to $Q^D(P_{\max}) - Q^*$. The implicit assumption underlying the conventional analysis is that only those demanders who *would* have wanted the good at the market-clearing price actually receive it. The imposition of a price ceiling brings new demanders into the market, however. These new demanders value the good less than the original demanders, but they are nevertheless likely to be allocated some portion of the limited supply: There is simply no incentive, in the presence of a price ceiling, for suppliers to identify and sell to the highest benefit demanders. The bottom-up deadweight-loss triangle measures the cost of this misallocation under the particular assumption of random rationing.

A second example in which explicitly considering bottom-up inefficiency significantly changes the way economists conceptualize and visualize the costs of market interventions is the case of quotas or other quantity restrictions. Unless a tradable and frictionless quota system is costlessly set up to do the secondary allocation, there will be a bottom-up inefficiency that is similar to that imposed by a price ceiling or a price floor. When one takes into account these bottom-up inefficiencies, the oft-maintained textbook equivalency of a tariff and a quantity restriction no longer holds. Quantity restrictions tend to be more costly than tariffs because the former induce bottom-up inefficiencies and the latter, typically, do not.

Bottom-up inefficiencies are also relevant—but typically neglected—in a third case: the analysis of monopoly and monopoly power. The standard approach focuses on the inefficiently low production chosen by producers with monopoly power, such as monopolists or oligopolists, that is, on the inefficiencies *conditional* on a given allocation of market power. This misses possibly important bottom-up inefficiencies that may result from the *allocation of* the "rights" to that restricted production quantity. The presence of monopoly power is indicative of some sort of (potentially unavoidable) pathology in the market. There therefore should be no presumption that the market has efficiently allocated the market power in the first place—indeed, the presumption should be that it has *not*. So, in addition to the top-down costs associated with an inefficiently low quantity supplied (the Harberger triangle), there are likely to be additional bottom-up costs associated with the *wrong* supplier(s) supplying the restricted quantity. That is, not only is the market inefficient *conditional on* the monopoly power, but it is also inefficient because the firms that are likely to have that power may not be the most efficient ones.

5. The Quantitative Importance of Bottom-Up Inefficiencies

If bottom-up inefficiencies were relatively small, then their underemphasis vis-à-vis top-down costs would be justifiable. This is not the case, however: For modest market distortions, the bottom-up costs are typically much *larger* than the top-down costs captured by the Harberger triangle.

To see the importance of these bottom-up costs, refer back to Figure 1 and the numerical example at the end of section 2. In Figure 1, the Harberger triangle EFC captures the top-down inefficiency, and the bottom-up inefficiency is captured by triangle EFA. These two triangles thus share a base EF, the length of which is given by the price gap ΔP between the demand and supply curves at the demand-restricted quantity $Q^D(P_{\text{min}})$. The height of the top-down triangle (relative to base EF) is $Q^* - Q^D(P_{\text{min}})$—that is, by the quantity distortion δ induced by the price floor. The height of the bottom-up triangle (relative to the same base) is simply $Q^D(P_{\text{min}})$, or, equivalently, by $Q^* - \delta$.

Because the top-down and bottom-up triangles share a base, the ratio of their areas is equal to the ratio of their heights. The ratio of the bottom-up to the top-down inefficiency is thus measured by $Q^D(P_{\text{min}})/(Q^* - Q^D(P_{\text{min}}))$; letting $\delta = Q^* - Q^D(P_{\text{min}})$ be the quantity distortion induced by the price floor, the same ratio can also be measured as $(Q^* - \delta)/\delta$. (Note that the validity of this formula depends only on the linearity of the *supply* curve.) In the numerical example at the end of section 2, the price floor reduced the quantity hired by $\delta = 10$ from the market-clearing level of $Q^* = 40$, and the ratio of the bottom-up to the top-down inefficiency was equal to $30/10 = 3$.

The top-down inefficiency—that is, the area of the Harberger triangle EFC in Figure 1—can, as usual, be written as $1/2(\Delta P(\delta))\delta$, where $\delta = Q^* - Q^D(P_{\text{min}})$ is the magnitude of the quantity distortion induced by the price floor, and $\Delta \text{P}(\delta)$ is the wedge between the supply and demand price at the distorted quantity (distance EF in Figure 1 or, more generally, $(P^D(Q^* - \delta) - P^S(Q^* - \delta))$. For linear supply curves, the ratio of the bottom-up to top-down inefficiencies is given by $(Q^* - \delta)/\delta$, so the bottom-up inefficiency is $1/2(\Delta P(\delta))(Q^* - \delta)$.[5]

Comparing the formulas $1/2(\Delta P(\delta))\delta$ and $1/2(\Delta P(\delta))\delta(Q^* - \delta)$ for the top-down and bottom-up inefficiencies caused by a price-ceiling-induced quantity distortion δ reveals two useful observations. First, estimating of the bottom-up inefficiency is no harder than estimating the "traditional" top-down inefficiency. The ratio of the bottom-up to the top-down costs is given by $(Q^* - \delta)/\delta$. To compute the bottom-up inefficiency, one therefore needs only two things: the top-down inefficiency and the magnitude (in percentage terms) of the distortion induced by the market intervention. Because the latter is necessary for computing the top-down inefficiency in the first place, computing the bottom-up inefficiency is as straightforward as computing the traditional Harberger inefficiency. For example, say that the restriction distorts quantity by 10% and that the top-down distortion measured by the Harberger triangle has been estimated to be $1 billion dollars. Then the bottom-up distortion would be nine times that or $9 billion.[6]

Second, it reveals that for small quantity distortions, the top-down inefficiency is dwarfed by the bottom-up inefficiency. In particular, the Harberger triangle is vanishingly small for small distortions; specifically, it is second order in δ for small δ. In contrast, the bottom-up triangle is *first* order in δ for small δ. For small distortions, traditional measures of deadweight loss completely miss the most important source of inefficiency.

Figure 3 plots the ratio of the bottom-up to the top-down costs as a function of the quantity distortion δ.[7] Notice that this ratio blows up as δ approaches zero, indicating the overwhelming importance of bottom-up inefficiencies for small distortions (which are the large majority of cases). Top-down distortions become more relevant than bottom-up distortions only for quantity distortions of over 50%! Because the Ramsey analysis behind the Harberger triangle was designed only to capture the consequences of small distortions, this suggests that the pedagogical weight given to the two types of inefficiencies should be exactly the reverse of what it is now, with bottom-up inefficiencies receiving substantially greater emphasis. A diagram simultaneously depicting both makes that possible.

6. Conclusion

The Harberger triangle is the wrong picture to use for depicting the welfare losses from market restrictions such as price or quantity controls. It fails to capture important bottom-up inefficiencies associated with these restrictions. The Harberger triangle for price controls implicitly assumes that somehow the produced (or consumed) units are allocated to the lowest-cost producers (or highest benefit consumers). As such, it tacitly frames the problem as if it were a decision to be made by a single agent akin to a monopolist—either a social planner or a

[5] This formula relies on the linearity of the supply curve, but qualitative conclusions clearly carry over to the more general case.

[6] In their analysis of the natural gas market, where they measured bottom-up inefficiencies directly rather than indirectly as we do, Davis and Kilian (2008) find that the bottom-up costs effectively triples the net welfare loss from gas price controls to consumers as compared to the loss measured by the Harberger triangle.

[7] Note, in particular, that this same ratio applies for any linear supply curve (for the case of a price ceiling, the same formula would apply for any linear demand curve).

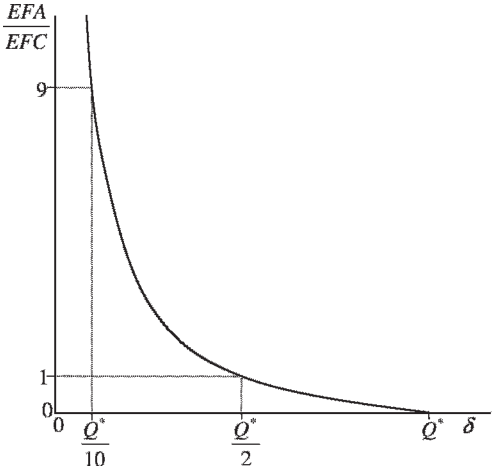

Figure 3. The ratio of the bottom-up to top-down costs from a quantity distortion δ. A binding price floor $P_{\min}$ distorts the quantity traded in a market by $\delta = Q^* - Q^D(P_{\min})$, where Q^* is the free market equilibrium quantity and $Q^D(P)$ is the demand curve. This graph plots the ratio of the bottom-up and top-down inefficiency costs as a function of the magnitude of this quantity distortion.

"representative agent." It thus obscures the key question: How does *the market*, made up as it is of a complex set of distinct and competing consumers and producers, actually allocate the production or consumption? In other words, the Harberger triangle approach is implicitly a top-down view of a phenomenon better seen from the bottom-up.

Paul Samuelson once said, "I don't care who writes a nation's laws or crafts its advanced treaties, if I can write its economics textbooks" (Nasar 1995, p. D1). His point was that what is in the texts matter, and because pictures are worth a thousand words, the illustrations we present to students matter greatly. By not providing students with a visual picture of the bottom-up inefficiencies that accompany price controls, and emphasizing only the top-down inefficiencies, we are providing a visual/discussion disconnect for the students and sending them out into the world with an underestimate of the costs of price controls and market restrictions. Our little picture helps remedy that.

Appendix 1

It is surprising that something as simple as what we are presenting in this article has not found its way into the standard texts. As we state in main body of the article, the general knowledge that there will be an allocative cost in addition to the Harberger triangle is well known—indeed, it is discussed in some popular introductory texts. What is not generally known is that there is an easy graphical way of capturing that cost. Once we came upon the method as part of work we were doing on another issue concerning the theory of price controls, we conducted a search of the literature to see if our graphical exposition had been developed elsewhere. We did not find anything, and people we shared the article with did not know of anywhere else it was to be found. However, happenchance led us to find previous expositions.

The happenchance occurred when, the day after Ted Bergstrom had commented upon our article, he attended a seminar by Davis and Kilian. In the paper they presented at that seminar, which was devoted to calculating an empirical measure of these bottom-up costs in the case of the natural gas markets, they presented a graph that was analogous to ours for the case of a price ceiling (Davis and Kilian 2008.) Ted e-mailed us the next day telling us "By sheer luck, we had a seminar today that bears directly on the paper you sent me. It does what seems to me an extremely nice job of quantifying the misallocation resulting from natural gas price ceilings (with the wrong houses getting access to gas)." We immediately looked at the Davis and Kilian paper and found a graph similar to ours. We also found that in their presentation, they referred to work by Paul MacAvoy and Robert Pindyck (1975) and Ronald Braeutigam (1981), where they had gotten the idea. Looking in these works we found the discussion of this idea for the case of natural gas (MacAvoy and Pindyck 1975, p. 54; Braeutigam 1981, pp. 161–3, with Braeutigam's being the most developed). So they deserve credit for coming up with the graph before we did.

That something as simple as this has been developed before is not surprising to us. What is surprising is that it has not been generalized and integrated into the texts, even by one of its early developers. It seems to be a case of a $100 graph lying on the sidewalk and no one picking it up, even those who dropped it! No one attempted to extend the graph beyond the discussion of the natural gas market, and thus it has not made its way into the textbooks or even into discussions of other market restrictions that were highlighting the importance of bottom-up inefficiencies, such as Glaeser and Luttmer (2003).

Our presentation differs from earlier presentations in the following ways. First, earlier presentations developed the idea only for price ceilings, whereas we generalize it for all quantity-based restrictions. Second, they motivate their analog of our "Effective Marginal Benefit" curve differently, presenting it as the demand curve for a "preexisting" customers; we present it as an explicitly constructed subset of a fixed set of demanders. Third, we explicitly determine the quantitative relationship between the Harberger top-down efficiency loss and the bottom-up efficiency loss, and we demonstrate that that relationship can be used to measure the bottom-up efficiency loss of quantitative restrictions.[8]

Given the quantitative importance of bottom-up or allocative inefficiency, as a pedagogical tool it would seem that this triangle should be given prominence in the texts. We hope that this article extends its use and makes the graph a key component of the textbook presentation of the costs of regulatory restrictions on markets.

References

Baumol, William, and Alan Blinder. 1998. *Economics: Principles and policies.* 7th edition. Fort Worth, TX: The Dryden Press.

Braeutigam, Ronald. 1981. The deregulation of natural gas. In *Case studies in regulation: Revolution and reform*, edited by Leonard Weiss and Michael Klass. Boston: Little, Brown and Company, pp. 142–86.

Davis, Lucas, and Lutz Kilian. 2008. The allocative cost of price ceilings in the U.S. residential market for natural gas. NBER Working Paper No. 14030.

Frank, Robert, and Ben Bernanke. 2009. *Principles of microeconomics*. 4th edition. New York: McGraw-Hill/Irwin.

Friedman, Milton, and George Stigler. 1946. Roofs or ceilings? The current housing problem. Reprinted in *Rent control, myth and realities: International evidence of the effects of rent control in six countries*, edited by Walter Block and Edgar Olsen. Vancouver, British Columbia: Fraser Institute, pp. 85–103.

Glaeser, Edward, and Erzo Luttmer. 2003. The misallocation of housing under rent control. *American Economic Review* 93(4):1027–46.

Hubbard, R. Glenn, and Anthony O'Brien. 2007. *Microeconomics.* 2nd edition. Upper Saddle River, NJ: Pearson Prentice Hall.

Kocieniewski, David. 2008. For Rangel, four rent-stabilized apartments. *New York Times*, 11 July, p. A1.

Krugman, Paul, and Robin Wells. 2005. *Microeconomics*. New York: Worth Publishers.

MacAvoy, Paul, and Robert Pindyck. 1975. *The economics of the natural gas shortage*. Amsterdam: North Holland Publishing Company.

Mankiw, Gregory. 1998. *Principles of economics*. Fort Worth, TX: The Dryden Press.

Nasar, Silvia. 1995. Hard act to follow? Here goes. *New York Times*, 14 March, p. D1.

Posner, Richard. 1975. The social costs of monopoly and regulation. *Journal of Political Economy* 83(4):807–28.

Tullock, Gordon. 1967. The welfare costs of tariffs, monopolies, and theft. *Western Economic Journal* 5(3):224–32.

[8] We also implicitly show that Davis and Kilian's claim that "the allocative cost only depends on the location and shape of the demand curve and the equilibrium level of price and quantities, but not on the shape of the supply curve" (Davis and Kilian 2008, p. 8) is misleading: The shape of the supply curve determines the constrained equilibrium quantity with price controls and thus plays a role in determining the amount of the allocative costs.