
# PIGOU, KNIGHT, DIMINISHING RETURNS, AND OPTIMAL PIGOUVIAN CONGESTION TOLLS


JOHN F. McDONALD



Journal of the History of Economic Thought / Volume 35 / Issue 03 / September 2013, pp 353 - 371
DOI: 10.1017/S1053837213000217, Published online: 30 July 2013

*Journal of the History of Economic Thought,*
*Volume 35, Number 3, September 2013*

# PIGOU, KNIGHT, DIMINISHING RETURNS, AND OPTIMAL PIGOUVIAN CONGESTION TOLLS

BY
JOHN F. McDONALD

*Arthur Pigou introduced the iconic two-road model in the first edition of* The Economics of Welfare *(1920), and it has been thought that this model was intended to demonstrate the need for Pigouvian taxes to mitigate traffic congestion. However, Pigou's intention was to show that efficient output for industries subject to decreasing returns required a tax on output. Pigou was incorrect, but the two-road model (correctly considered) became the starting point for the analysis of traffic congestion in the 1950s. This paper recounts the doctrinal history of decreasing returns industries and the two-road model.*

## I. INTRODUCTION

This paper recounts, places in historic context, and interprets the controversy between A. C. Pigou, Allyn Young, and Frank Knight regarding diminishing returns in competitive industries. A model introduced by Pigou (1920) and explained in detail by Knight (1924) as part of that debate was, three decades later, employed by transportation economists. This paper examines the reintroduction of this model, which is a cornerstone of modern transportation economics. Pigou (1920) is considered to be the originator of the idea that congestion tolls should be used to achieve efficient use of a road that suffers from the negative external effect of traffic congestion. For example, Santos and Verhoef (2011, p. 561), in a recent survey chapter, state:

Roosevelt University – Real Estate, Chicago, IL. Email: mcdonald@uic.edu. The author thanks an anonymous reviewer for extensive helpful comments on earlier versions of the paper. The paper was presented at the annual North American Meetings of the Regional Science Association International on November 10, 2011. The author thanks discussant Jonathan Hall and session attendees for their comments.

ISSN 1053-8372 print; ISSN 1469-9656 online/13/03000353-371 
doi:10.1017/S1053837213000217

> Road pricing has long been viewed as a potentially efficient instrument for dealing with traffic congestion. In 1920 Arthur Pigou used the example of a congested road to explain the economics of external effects, and in particular how a corrective tax can be used to restore efficiency when some goods are not optimally priced at marginal cost.

Pigou did present the example to which Santos and Verhoef refer, but it was not intended as a case of external effects in which, to use the terms employed by Pigou, private and social marginal net products diverge. The purpose of the example was entirely different. Since taxes to correct inefficiencies caused by negative external effects are called "Pigouvian taxes," it is natural to think that, in 1920, Pigou intended to apply the concept to traffic congestion. Such was not the case.

Pigou (1912) asserted that a competitive industry faced with rising costs expands output beyond the efficient level. Later, he presented in the first edition of *The Economics of Welfare* (1920, p. 194) a model of two roads—one less troublesome (uncongested) and one more troublesome (congested)—as an example of diminishing returns in a competitive industry, and argued that the industry subject to diminishing returns (the congested road) expands output beyond the efficient level. Young (1913) and Knight (1924) argued successfully that Pigou's general proposition is incorrect, and Knight showed that the two-road example does not demonstrate Pigou's proposition. Pigou (1924) dropped the model of the two roads from the second edition and subsequent editions of *The Economics of Welfare*, and eventually acknowledged that a competitive industry subject to "diminishing returns" does not expand output beyond the efficient level in the absence of negative externalities. The fourth edition of *The Economics of Welfare* (1932), a text that contains no discussion of congested roads, was studied by generations of students.

According to Knight (1924), the fallacy in Pigou's analysis is the failure to consider what Knight called "entrepreneur's cost." Knight showed that, if the more troublesome road were owned by a private firm, the firm would set a price for the use of the road that is efficient. In effect, Knight anticipated the Coase theorem that the proper definition of property rights can produce the efficient allocation of resources. A corrective tax was unnecessary.

However, the model of the two roads came back to life in the 1950s, perhaps because the article by Knight (1924), in which he carefully explains the model, was reprinted in *Readings in Price Theory* (1952), edited by Stigler and Boulding. This volume was used widely in graduate economics programs in the 1950s and 1960s.[1] A. A. Walters (1954) appears to be the first transportation economist to refer to the two-road model in a discussion of congestion tolls on roadways. The Pigou–Knight model was used by other transportation economists as the starting point for analyses of the efficient use of existing roadways and investment in new roads. A model that initially was applied incorrectly turned out to be highly useful in another application, but the newer, correct application did not occur until thirty years later.

This paper first discusses the context of economic thought in which the Pigou–Young–Knight debate took place, and then examines Pigou's original arguments and

[1]The article by Knight (1924) is the only article that appears in both *Readings in Price Theory* (1952) and in *Readings in Welfare Economics* (1969), edited by Kenneth Arrow and Tibor Scitovsky. It is included in the section on costs and returns in the earlier volume, and in the section on social versus private costs and benefits in the later volume.

the critiques by Young, Knight, and others. Emphasis is placed on the final resolution of the controversy regarding "diminishing returns" in competitive industries by Ellis and Fellner (1943). Then, the rebirth of the two-road model and its use by Walters (1954); Beckmann, McGuire, and Winsten (1956); and others is discussed.

## II. MARSHALL AND PIGOU

Two of the most influential texts of twentieth-century economics were published in 1920. Alfred Marshall produced the eighth and final edition of *Principles of Economics*, and A. C. Pigou published the first edition of *The Economics of Welfare*—an updated version of his earlier *Wealth and Welfare* (1912). Both texts were published by Macmillan & Co. and, of course, both Marshall and Pigou were at Cambridge and can be regarded as the most important figures in English economic thought of that time.

Marshall's *Principles* had been in widespread use since the first edition was issued in 1890. Marshallian economics is the context in which Pigou functioned at this time. It is fair to say that Marshall saw long-run continuous progress in both economic science and the ability of the market economy to provide for the "the material requisites of wellbeing" (1961, p. 1). The preface to the first edition, which was reprinted in subsequent editions, introduced the Principle of Continuity and states (1961, p. vii):

> The notion of continuity with regard to development is common to all modern schools of economic thought, whether the chief influences acting on them are those of biology, as represented in the writings of Herbert Spencer; or of history and philosophy, as represented by Hegel's *Philosophy of History*, and by more recent ethico-historical studies on the Continent and elsewhere. These two kinds of influences have affected, more than any other, the substance of the views expressed in the present book; but their form has been most affected by mathematical conceptions of continuity.…

Marshall's *Principles* involved explicating the laws of demand and supply in the context of a market economy that was changing continuously. Consider briefly the organization of the text. After two introductory books, Book III is titled *On Wants and Their Satisfaction*, introduces basic concepts such as marginal utility and consumer's surplus, and discusses how demands change and grow over time.

Book IV is a lengthy book that examines "The Agents of Production." The final chapter in Book IV (1961, p. 262) is titled "Conclusion. Correlation of the Tendencies to Increasing and Diminishing Return." Briefly stated, the big question addressed in Book IV is whether (and under what conditions) the expansion of an industry in the long run is accompanied by increasing, constant, or diminishing returns at the industry level. Marshall stated (1961, p. 262):

> At the beginning of the Book we saw how the extra return of raw produce which nature affords to an increased application of capital and labour, other things being equal, tends in the long to diminish. In the remainder of the Book and especially in the last four chapters we have looked at the other side of the shield, and seen how man's power of productive work increases with the volume of work that he does.

Marshall discussed in detail how increasing returns to an industry can arise from both internal and external economies. The book includes the famous Chapter X on the "Concentration of Specialized Industries in Particular Locations" that includes the statement that (1961, p. 225):

> When an industry has thus chosen a locality for itself, it is likely to stay there long: so great that the advantages which people following the same skilled trade get from near their neighborhood to one another. The mysteries of the trade become no mysteries; but are as in the air, and children learn many of them unconsciously.

Book V shows the implications of books III and IV for the workings of markets; demand, supply, marginal cost, and value in the long and short periods. Marshall concludes in the case of increasing returns (1961, p. 391):

> In the case then of commodities with regard to which the law of increasing return acts at all sharply, or in other words, for which the normal supply price diminishes rapidly as the amount produced increases, the direct expense of a bounty sufficient to call forth a greatly increased supply at a much lower price, would be much less than the consequent increase in consumers' surplus. And if a general agreement could be obtained among consumers, terms might be arranged which would make such action remunerative to the producers, at the same time that they left a large balance of advantage for consumers.

Such a general agreement would enhance the ability of society to benefit from the continuous change that Marshall saw taking place. He goes on to state:

> One simple plan would be the levying of a tax by the community on their own incomes, or on the production of goods which obey the law of diminishing returns, and devoting the tax to a bounty on the production of those goods with regard to which the law of increasing return acts sharply. (p. 392)

Note that Marshall suggested a tax on goods that obey the law of diminishing returns in conjunction with a bounty for goods subject to increasing returns. Such a tax and bounty system might be of net benefit to society—the loss of consumers' surplus with the tax might be more than offset by the gain of consumers' surplus from the bounty. Also, note that Marshall did not conclude that an industry producing under conditions of diminishing returns will produce more than the ideal output for that industry. Book VI, the final book, examines the functional distribution of income that results from the workings of the market economy.

Enter Pigou. His topic in *The Economics of Welfare* (and in the earlier *Wealth and Welfare*) is the welfare of society as measured by the national dividend, expressed as "everything that people buy with money income together with the services that a man obtains from a house owned and inhabited by himself" (1961, p. 34). This follows the precedent set by Marshall in the *Principles*. Part I of *The Economics of Welfare* discusses the national dividend, its measurement, its distribution, and its relationship to the welfare of society. Part II is a lengthy presentation of the theory of resource allocation, with numerous examples. Pigou stated (1961, p. 127):

> In this Part we are concerned with causes that increase or diminish the size of the national dividend by acting on the way in which the productive resources of no matter what kind belonging to the country are distributed among different uses or occupations.

At the outset, Pigou stated that the problem is more complex than simply ensuring that permitting people to act in their own self-interest will promote the welfare of society as a whole. He quoted Marshall (1961, p. 394):

> Much remains to be done, by a careful collection of statistics of demand and supply and a scientific interpretation of their results, in order to discover what are the limits of the work that society can with advantage do towards turning the economic actions of individuals into those channels in which they will add the most to the sum total of happiness.

The distinction between marginal social and private net products is fundamental to Pigou's analysis. His definitions (1961, pp. 134–135) are as follows.

> The marginal social net product is the total net product of physical things or objective services due to the marginal increment of resources in any given use or place, no matter to whom any part of this product may accrue.
>
> …
>
> The marginal private net product is that part of the total net product of physical things or objective services due to the marginal increment of resources in any given use or place which accrues in the first instance—i.e., prior to sale—to the person responsible for investing resources there.

Part II includes numerous examples of divergences between marginal social net product and marginal private net product, including obstacles to movement, imperfect knowledge, state regulation, industrial forms (e.g., monopoly), and three classes of divergences that can occur under unhindered normal competition. One case involves persons who are not producers of the commodity in question (p. 183):

> Here the essence of the matter is that one person A, in the course of rendering some service, for which payment is made, to a second person B, incidentally also renders services or disservices to other persons (not producers of like services) of such a sort that payment cannot be exacted from the benefitted parties or compensation enforced on behalf of injured parties.

The second case concerns services or disservices that accrue to other producers of the commodity in question. The third case involves owners of durable instruments of production that are rented by the producer. It so happens that the distinction between the first and second cases is critical to the understanding of the Pigou–Young–Knight controversy, which involves the second case and Pigou's analysis of industries subject to increasing or decreasing returns in the sense of Marshall. Pigou concluded in *Wealth and Welfare* (1912) that industries subject to increasing returns, in order to produce the ideal output, should receive a bounty, as Marshall had stated. He also concluded that industries subject to decreasing returns should be taxed because otherwise output will be larger than ideal, even in the absence of disservices imposed on other producers for which no compensation is paid. Pigou seemingly had supplied Marshall's suggestion for a system of taxes and bounties with stronger theoretical support. The story of the Pigou–Young–Knight controversy begins here.[2]

[2]The controversy includes the "empty boxes" controversy that appeared in the *Economic Journal* during 1922 to 1924. J. H. Clapham (1922) questioned the usefulness of the concepts of diminishing, constant, and increasing return industries; Pigou (1922) defended them; and Robertson (1924, [with reference to Young [1913] and others] showed that the diminishing and increasing returns cases are not parallel.

## III. *WEALTH AND WELFARE*

The story of the Pigou–Young–Knight controversy begins with Chapter VIII in Pigou's *Wealth and Welfare*. In this chapter, he considers a competitive industry, and begins by stating (1912, p. 172), "Let it be assumed that private supply prices and social supply prices coalesce throughout." He then refers to a diagram reproduced as Figure 1. Supply curve $SS_1$ is a supply curve of the "ordinary type" and supply curve $SS_2$ is called (1912, p. 173) the "curve of marginal supply prices $SS_2$ such that (it) represents the difference made to the aggregate expenses of the industry concerned with the production of the (marginal) unit of output." Pigou argued that the industry will produce the level of output that is determined by the intersection of the demand curve and the supply curve of the "ordinary type," but that the ideal output is determined by the intersection of the demand curve and the "curve of marginal supply prices $SS_2$." In Figure 1, actual output exceeds ideal output.

Allyn Young (1913) recognized Pigou's mistake immediately in a review of *Wealth and Welfare*. Referring to the curve that Pigou called the "difference made to aggregate expenses," Young (1913, p. 681) stated: "But I fail to see that its use is appropriate in the analysis of the extent to which competition tends to secure the maximum national dividend." Young went on to state (1913, p. 683): "Increased prices for the use of land and the other factors of production do not represent an increased using up of resources in the work of production. They merely represent transferences of purchasing power." As Ellis and Fellner (1943, p. 498) stated, Young could have expressed the point as, "If the expansion of an industry gives a factor a higher per unit remuneration, whether or not that higher price induces a greater aggregate supply of the factor, the units already being supplied earn producers' rents (or increase in the previous rent); and rent is not a cost in social resources." Actually, Young (1913, p. 677) did use the word "rent."

There are three standard reasons for a divergence between the supply curve of the ordinary type and the curve for the difference make to aggregate expenses: externalities, rising supply price on one or more inputs; and diminishing returns to scale at the *industry* level. Pigou ruled out the first of these by assumption in this particular instance, and Young (1913) based his argument primarily on the second reason, but included reference to the third reason. Young stated (1913, p. 676):

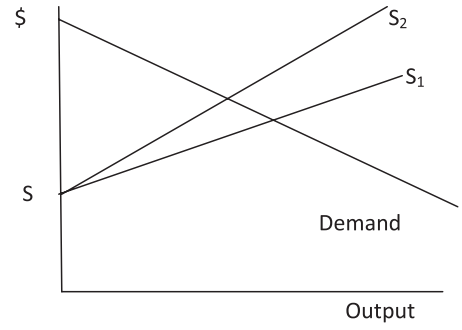


FIGURE 1. Pigou's Supply Curves.

> The facts to be observed are those connected with the increase in the aggregate expenses of an industry as a whole. These must be distinguished from the general fact of the diminishing productivity of particular factors of production and also from the tendency to diminishing (or constant, or increasing) returns per unit of expense as the size of the individual business unit increases. Furthermore, we do not have in mind precisely the same thing as when we speak of diminishing (or constant, or increasing) returns in a given industry as a whole during a period of time, although the basic facts involved in both conceptions are in part identical. Take, for example, an industry of diminishing returns, say wheat growing. By 'diminishing returns' we mean, of course, to imply that if an aggregate annual product of x units of wheat is increased to an annual product of $x + \Delta x$ units, more capital and labor per unit of product must be 'applied to the land' to produce the $\Delta x$ units than were required to produce the final increments of the original x units.
>
> Now that fact of present importance is that, through the rise in land rent (and possibly, also, through an increased expense per unit of labor and capital, caused by the increased demand) the aggregate expenses, including land rent, of producing $x + \Delta x$ units will exceed the expenses of producing x units by much more than the expenses specifically attributed to the production of the $\Delta x$ units.

Pigou responded to Young in the first edition of *The Economics of Welfare* in 1920.

## IV. *THE ECONOMICS OF WELFARE*, FIRST EDITION

In *The Economics of Welfare* (1920), Pigou agreed that Young's argument was correct for rising transfer prices, but not for the case of diminishing returns to the inputs. Pigou (1920) continued to maintain that competitive industries subject to diminishing returns at the industry level produce an inefficiently large level of output. He illustrated the point using the example of the two roads. His entire statement of the two-road example is as follows (1920, p. 194):

> Suppose there are two roads ABD and ACD both leading from A to D. If left to itself, traffic would be so distributed that the trouble involved in driving a 'representative' cart along each of the two roads would be equal. But, in some circumstances, it would be possible, by shifting a few carts from route B to route C, greatly to lessen the trouble of driving those still left on B, while only slightly increasing the trouble of driving along C. In these circumstances a rightly chosen measure of differential taxation against road B would create an 'artificial' situation superior to the 'natural' one. But the measure of differentiation must be rightly chosen.

Read in isolation, there is nothing incorrect in this statement. The problem is that this example appears in a discussion of "simple competition" in the absence of divergences between "marginal social net product and marginal private net product" (i.e., externalities). The sentences immediately preceding the two-road example are (1920, p. 194):

> This result is of considerable theoretical importance, because it is in direct conflict with the widespread opinion that, apart from certain possible indirect effects, differential taxes are necessarily wasteful and necessarily cause people to obtain what they want

> by a more costly, instead of a less costly, route. This opinion is incorrect, and the nature of the error can be easily illustrated.

Pigou (1912, pp. 148–171; and 1920, pp. 149–179) provided numerous examples of divergences between marginal social and private net product that pertain to persons who are not producers of like goods and services, but did not include traffic congestion as an example. Perhaps the most famous example is (1920, p. 160): "It has been said that in London, owing to the smoke, there is only 12 per cent as much sunlight as is astronomically possible, and that one fog in five is directly caused by smoke alone, while all the fogs are befouled and prolonged by it."

Pigou provided a lengthy justification for his original position that a competitive industry subject to diminishing returns produces more than an efficient level of output (1920, appendix 3). He acknowledged that Professor Young was correct in the case of rising input supply prices, but dismissed this case on the grounds of limited relevance because variations in output produced by an individual industry are unlikely to have appreciable effects on input prices. Instead, he stated that (1920, p. 936):

> The reason why diminishing returns in terms of money appear is, in general, not that the money price of factors employed is increased, but that that proportionate combination of different factors, which is most economical to employ when $(x + \Delta x)$ units of commodities are being produced is a less efficient proportionate combination than that which it is most economical to employ when x units are being produced; and the extra cost involved in this fact is real, not merely nominal. For these reasons Professor Young's objection, as a general objection, fails.

However, this lengthy appendix does not include mention of the two-road example. Apparently, Pigou thought of the two roads as different industries, and thought that cart traffic was the output of each industry. In this case, the "industry" represented by route B is producing an inefficiently large output (cart traffic) and, therefore, a tax should be imposed on output. But what are the firms in this industry? Each cart is (perhaps) a "firm" that produces cart trips and somehow there is more "trouble" on route B than on route C. Note that Pigou does not give this "trouble" a name.

## V. KNIGHT'S REFORMULATION OF THE TWO-ROAD MODEL

Pigou (1920) had developed (very briefly) the two-road model to demonstrate his contention that a competitive industry subject to diminishing returns produces a quantity of output that is greater than the efficient amount, but Knight (1924) provided a complete presentation of the model. There are two roads that connect an origin and a destination, and the roads are used by trucks (i.e., competitive firms). One road is wide and never subject to increasing costs, but is poorly surfaced. The other road is of much better quality, but is narrow and has a limited capacity. The speed limit is higher on the second road, but actual speed falls as the number of users increases. Pigou (1920, p. 194) argued that "carts" will use both roads, and will utilize the second road up to the point at which the driving time on the second road equals the (given) driving time

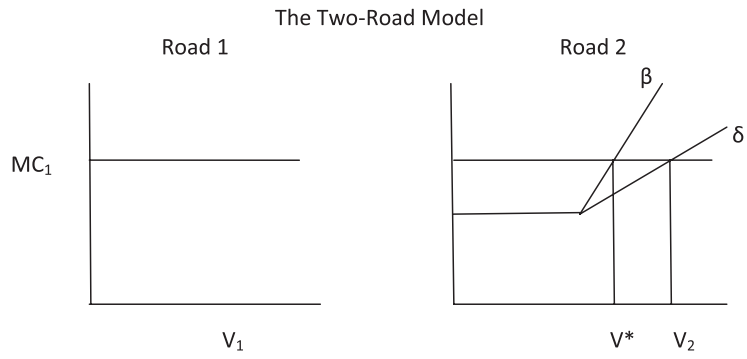


FIGURE 2. The Two-Road Model.

on the first road. Knight switched the vehicles to trucks.[3] This solution is inefficient because inducing the marginal truck to switch from the second road to the first road will save time for the trucks that continue to use the second road, while no added cost is imposed on the truck that switches to the first road. Pigou concluded that the efficient solution is to impose a differential tax on vehicles that use the second road (i.e., the output of the second "industry" is inefficiently large).

Knight's version of Pigou's argument is illustrated in Figure 2. The average and marginal cost function for trucks that use Road 1 is a constant $MC_1$, while the cost functions for Road 2 are labeled δ and β. Cost function β is the function that represents the increase in total cost incurred by producing one more unit (one truck trip). Professor Knight's presentation of Pigou's model asserted that individual trucks base their decisions on cost function δ, and, therefore, too many trucks use Road 2. At usage level $V_2$, the cost of the last truck trip on Road 2 (on the β function) far exceeds the cost of a trip on Road 1. The efficient level of usage is V*.

The question is: what are cost functions δ and β? Knight (1924, p. 585) provided the answer. He stated:

> As more trucks use the narrower and better road, congestion develops, until at a certain point it becomes equally profitable to use the broader but poorer highway. The congestion and interference resulting from the addition of any particular truck to the stream of traffic on the narrow but good road affects in the same way the cost and output of all of the trucks using that road.

The critical assumption in the model involves congestion and interference (of trucks with each other) on Road 2. Cost function δ is the cost as seen by each trucker: the average cost of a trip. Cost function β is the cost of an additional truck trip for the trucks as a group, including the congestion cost. Pigou rightly saw that truckers respond to the average trip cost, and ignore the marginal cost that includes the congestion cost. However, obviously this is a case of an external cost that is not imposed on the individual marginal trucker. It is not a case of diminishing returns to an

[3]Can any importance be attached to the change from carts to trucks? Knight lived in a large city (Chicago) that was subject to a great deal of traffic congestion at a time when the use of trucks was growing rapidly.

industry, as illustrated in the previous sections. Pigou's example did not pertain to the point he was trying to make. But Pigou's example did address an important problem, as is discussed below.

Knight (1924) further demonstrated that, if Road 2 is privately owned, the owner will set a toll equal the amount needed to produce efficient levels of traffic. Knight (1924, p. 587) states, "This is clearer if we think of the owner of the road hiring the trucks instead of their hiring the use of the road." This permits us to think of the cost of truck travel on the uncongested road as being equal to the value of the marginal product of a truck trip. Knight (1924, p. 587) then states, "The toll or rent will be so adjusted that the added product of the last truck which uses the narrow (congested) road is just equal to what it could produce on the broad (uncongested) road." The amount that a truck can produce on the congested road includes, of course, the amount that its presence on the road subtracts from the amounts produced by the other trucks on the road.

This point perhaps is demonstrated more clearly with a basic mathematical model. Assume that the total benefits of truck trips can be written B(V), and that the constant cost of truck trips on Road 1 is $C_1$. The cost of a trip for a trucker on Road 2 is $C_2(V_2)$, a function of the traffic volume. The problem is to maximize the net benefits of truck trips, written

$$NB = B(V) - C_1V_1 - C_2(V_2)V_2, \text{ where } V = V_1 + V_2. \tag{1}$$

Maximization of net benefits with respect to $V_1$ and $V_2$ produces the first-order conditions

$$B' = C_1 \text{ and}$$
$$B' = C_2 + V_2(dC_2/dV_2). \tag{2}$$

Here, B' is the marginal benefit of a truck trip and $V_2(dC_2/dV_2)$ is the marginal congestion cost imposed on all the users by the marginal truck. Therefore, efficiency requires that

$$C_1 = C_2 + V_2(dC_2/dV_2). \tag{3}$$

The efficient toll equals the marginal congestion cost.

Knight (1924) considered the profit-maximizing solution for an owner of Road 2. The demand for trips on this road is perfectly elastic at price $P = C_1$ because any trucker can take the uncongested road for this cost. The problem (in the short run) is to maximize toll revenue $\Pi$, where toll revenue is the difference between the total amount paid by all the trucks and the real cost of the trips;

$$\Pi = PV_2 - C_2(V_2)V_2 = C_1V_2 - C_2(V_2)V_2. \tag{4}$$

Maximization with respect to $V_2$ yields

$$d\Pi/dV_2 = C_1 - C_2 - V_2(dC_2/dV_2) = 0. \tag{5}$$

The toll charged by the owner of Road 2 is $V_2(dC_2/dV_2)$.

Clearly, Knight had made a convincing point. Pigou dropped the two-road model from the second and subsequent editions of *The Economics of Welfare*, and others

dismissed the two-road model. For example, Ellis and Fellner (1943) provided a complete restatement of the Pigou–Knight controversy, and stated that (1943, p. 501):

> Pigou's contention that of two roads connecting the same two points the one, assumed to be superior but narrow and therefore subject to diminishing returns, is overexploited in competition unless taxed differentially seems to have rested on the notion that competitive output is determined by the δ function. The contention was proven to be fallacious by Professor Knight, who has shown that the owner of the good road will charge a toll that will raise costs to users to the β level.… No special significance should be attributed to Pigou's recantation of the "two roads" proposition because it was inconsistent with his own position and should have been dropped even if his position had not been modified.

Evidently, the two-road model was dropped from the discussion because it was not relevant to the main issue of the efficiency of perfect competition in the presence of diminishing (or increasing) returns. Pigou had (perhaps inadvertently) put his finger on an important issue because most roads are subject to congestion and are not privately owned toll roads. But the importance of this case would not become evident for thirty years.

## VI. THE DEBATE CONTINUED: THREE COST FUNCTIONS

The debate over the decreasing and increasing returns cases continued until Ellis and Fellner (1943) provided the definitive analysis. This section refers to the contributions of Joan Robinson (1933, 1941), but otherwise omits references to the literature that was published after 1924. Ellis and Fellner began with the following (1943, p. 493):

> Along with its answer to the principal problem which it set for itself as to how competition allocates resources amongst various uses, neo-classical economics bequeathed to the present generation the much debated proposition that competition causes output under 'diminishing returns' to exceed, and under 'increasing returns' to fall short of, an output corresponding to the social optimum.

Some clarity in the matter of efficient output of competitive industries in the presence of increasing costs is obtained by displaying three cost functions. This presentation draws upon Ellis and Fellner (1943), who gave credit to Joan Robinson (1933, ch. 10). Increasing costs arise from inputs with rising supply prices and from diminishing returns to the inputs.[4] Simple mathematical models are provided in the appendix to demonstrate the basic results.

[4]Robinson (1941) developed a simple general equilibrium model to explain rising supply price. Her conclusion (1941, p. 5) is: "Thus, for any commodity considered separately, there is rising supply price, because an increase in the output of any commodity turns relative factor prices against itself." This tends to occur because, with full employment, the price of the input employed in relatively large amount (compared to the rest of the economy) is bid up, while the price of the input that is used in relatively small amount is bid down.

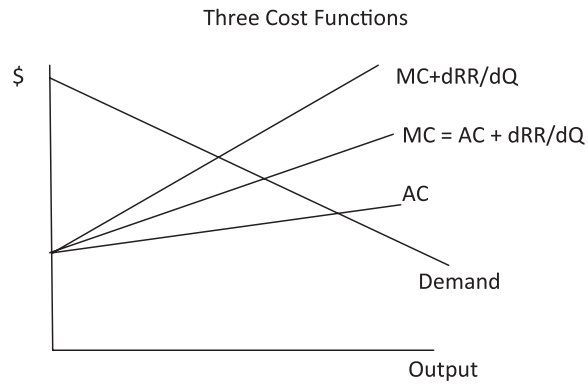


FIGURE 3. Three Cost Functions

The three cost functions are shown in Figure 3.[5] The cost function with the least upward slope (AC) is the standard average cost function for the industry, and it reflects rising input prices or diminishing returns (or both). The marginal cost function (MC) is the marginal cost for the industry, and equals average cost plus the change in Ricardian rent (RR) as output increases. Demonstrations of MC = AC + dRR/dQ are provided in the appendix. The steepest cost function is marginal cost plus the change in Ricardian rent, MC + dRR/dQ.

In *Wealth and Welfare,* Pigou (1912, p. 173) asserted that competitive industry output will be established at the intersection of the demand function and the marginal cost function (MC), but that the intersection of the demand function with marginal cost plus the increment of Ricardian rent determines the efficient allocation of resources. He thus concluded that a competitive industry subject to rising costs should be subject to a tax.

Pigou (1924, p. 31) admitted that his argument in the first edition of *The Economics of Welfare* was not adequate. His response to D. H. Robertson (1924) is:

> Professor Allyn Young's criticism of my analysis of diminishing returns, which Mr. Robertson has transcribed, is, in my present judgment, substantially valid as regards long-period problems, and the reply which I made to it in *The Economics of Welfare* is not adequate. In view of that criticism important modifications in my analysis are necessary, and are made in the forthcoming new edition.

As Ellis and Fellner stated (1943, p. 501):

> The revision consisted in the abandonment of the general thesis that, under increasing cost, output under competition exceeds the ideal, and the adoption of the very limited proposition that a divergence occurs only from the viewpoint of one nation against another when it pays agricultural rents to foreign owners in the price of imports.

The "very limited proposition" was stated in the second and third editions of *The Economics of Welfare*, but does not appear in the fourth edition. Ellis and Fellner (1943, pp. 503–504) showed that the limited proposition "has little or no validity." The particular case of agricultural rents paid to foreign owners is not under consideration here.

[5]Figure 2 is a reproduction of a diagram from Ellis and Fellner (1943), but uses different notation because their notation is confusing.

## VII. REVIVAL OF THE TWO-ROAD MODEL

The late A. A. Walters published his well-known article "The Theory and Measurement of Private and Social Cost of Highway Congestion" in *Econometrica* in 1961. In that article, he cited its theoretical antecedents: Pigou (1920); Knight (1924); Beckman, McGuire, and Winsten (1956); and Walters (1954). Based on this evidence, one can conclude that Walters can be credited with reviving Pigou's two-road model after thirty years. The article by Walters (1954) is a survey of policy issues surrounding "track costs and motor taxation." He considered various ideas regarding motor taxation that were current, and then added (1954, pp. 142–143):

> Another criterion, and I think a more substantial one, may be developed on the basis of the distinction between marginal private and marginal social cost. In the first edition of *Wealth and Welfare*, Professor Pigou produced the famous example of two roads connecting two points. One of those roads is assumed to be wide enough to accommodate all the traffic which might possibly use it, but it has a bad alignment and surface. The other road has a much better alignment and surface, but it is narrow and after a certain level of traffic flow there is some congestion. The additional private cost of a vehicle travelling over the narrow road will be reflected in the accounts of the firm or person operating the vehicle. If the vehicles are owned individually by a large number of firms and if no charge is made for the services of the road, the traffic level will be determined by the condition that marginal private cost equals price. But when the road is congested, an additional vehicle increases the costs of vehicles already using the road because congestion is increased. The additional costs of intra-marginal vehicles are not reflected in the accounts of the marginal vehicle, but, if factor prices are constant, they must be classified as part of the marginal social cost of transport on the narrow road. In these conditions, the number of vehicles using the narrow road will be greater than is socially desirable. It is argued that a tax should be imposed on vehicles using the narrow road so that the full marginal social costs are reflected in the accounts of individual vehicle operators. In modern terminology the difference between private cost and social cost would be called external diseconomies.[6]

Walters (1954) went on to suggest that the ideal tax is complicated and would vary from road to road and by time of day. And then he suggested (1954, p. 143):

> Nevertheless, it is clear that measures such as an annual license fee for vehicles using the central areas of large towns would be a move in the right direction. For example, a special 'London license' would have to be acquired and displayed before vehicles could use the roads of central London between 8 a.m. and 6 p.m.

This idea was implemented virtually in this exact form almost fifty years later (in 2003). Leape (2006) provides a good discussion of the London congestion charge.

---

[6]Note that Walters (1954) stated incorrectly that the reference for the two-road model is *Wealth and Welfare* (1912) rather than *The Economics of Welfare* (1920).

## VIII. THE TWO-ROAD MODEL IN TRANSPORTATION ECONOMICS

The two-road model is a crucial stepping stone to the construction of realistic models that can be used to examine important questions in transportation economics. The two early important contributions in the field by Beckmann, McGuire, and Winsten (1956), and Walters (1961), began with the two-road model. Two policy issues have been the focus of research in recent years: the effects adding to the capacity of a congested highway, and the design of an efficient congestion toll; and the two-road model provided important insights into both issues.

Downs (1962) introduced what he called the "law of peak-hour expressway congestion" in which highway capacity expansion is associated with an increase in demand that leaves congestion unchanged. The issue now is known as the "Pigou–Knight–Downs Paradox," as discussed by Arnott and Small (1994) and many others, and asserts that the expansion of the capacity of a congested highway can actually result in no reduction in congestion. An explanation for this "law" can be found in the two-road model shown in Figure 2. Suppose that the capacity of the congested road is expanded—but not to the extent that the road can carry all traffic demand at a cost less than the cost on the uncongested road. Traffic will shift from the uncongested road to the congested road until the cost of travel on the congested road returns to its former level equal to the cost of travel on the uncongested road. This result occurs as long as anyone is using the uncongested road. As noted above, it is assumed that the two roads are substitutes. The crucial assumption is that there is congestion on only one of the two roads. However, if both roads are congested, then an expansion of capacity will reduce the level of congestion and travel cost as long as the demand for travel is not of infinite elasticity.

Expansion of the model to include travel at other times of day and other modes of travel leads to what Downs (1962) called the "triple convergence." Expansion of the congested highway induces travelers to switch from the uncongested road, other times of day, and other modes of travel. The level of congestion during the "rush hour" does not decline, but positive results are the shortening of the rush hour and a possible reduction in congestion in time periods adjacent to the rush hour. Two other possible demand potential factors can be added: people who did not travel before but decide to travel after expansion of capacity, and people who change residential or employment locations in response to capacity expansion. This demand for additional trips on the expanded highway is now known as the "latent demand," and there is a growing body of empirical evidence demonstrating that latent demand is substantial in magnitude. See Cervero (2002) for a survey of the literature. The study by Duranton and Turner (2011) is a recent example.

As Walters (1954, 1961) discussed, a basic lesson from the two-road model pertains to the pricing of the congested road. The efficient congestion toll on the congested road equals the marginal congestion cost. The inclusion of the uncongested road in the model is essential because traffic diverted from the congested road by the imposition of a toll does not cause congestion elsewhere in the road system. The landmark volume by Beckmann, McGuire, and Winsten (1956) begins with a detailed explanation of the two-road model, which is followed by a greatly expanded model that includes an entire transportation network with many origins, destinations, and possible routes for

each origin–destination pair. Some links in the network may carry traffic with different origins, destinations, or both. The case of "flexible demand" (i.e., the demand for trips from origin i to destination j is a function of the cost of the trip) was examined (1956, pp. 92–94). The efficient utilization of the road network is achieved by imposing a congestion toll on each congested link in the network. As Beckmann, McGuire, and Winsten (1956, p. 94) stated:

> These tolls express the excess of social over private cost caused by an additional road user. The 'toll' term added to the private cost of transportation equals the increase in the private cost to the average road user caused by a unit increase in traffic, multiplied by the total flow of traffic.

In short, the overall efficient use of a road network (first-best solution) with many congested links requires tolls that are the same as in Pigou's simple two-road example. An uncongested link requires no toll. If both roads in Pigou's example are subject to congestion, then two tolls are needed to achieve efficiency.[7]

Beckmann, McGuire, and Winsten recognized the impracticality of charging tolls on every congested road, and briefly considered the problem of setting tolls in a world with only a limited number of toll roads.[8] They reached this prescient conclusion (1956, p. 97):

> This limited efficiency problem can be approached in terms of the maximization of our consumers' surplus function subject to the additional constraints that on each free road the difference in trips costs between two end points of the road should not exceed the average transportation cost on the road. Solution of this modified problem shows that best tolls on the toll roads are different from what they would be for the same roads in a general toll-road system in which efficiency tolls are charged on each congested road. They are higher where congested free roads are predominant as feeder roads, and are less where free roads compete with a toll road as alternatives.

However, Beckmann, McGuire, and Winsten (1956) did not provide analytical results for this conclusion. Clear analytical results were obtained by researchers who returned to the two-road model.

McDonald (1995) considered the problem in which only one road in the two-road model has a toll, and derived the following (second-best) toll:

$$\text{Toll} = V_t\left(dC_t/dV_t\right) + \left(dV_f/dV_t\right)V_f\left(dC_f/dV_f\right). \tag{6}$$

[7]McDonald (1995) showed that only one toll is required in the case of two roads if the total demand for trips is fixed.

[8]The impracticality of charging tolls on every congested road arises from both technical difficulties and resistance on the part of citizens. Computer and GPS technologies have reduced substantially the technical difficulties, but citizen resistance remains strong because the reduction in traffic congestion requires that the sum of time and monetary cost of using a congested facility must increase.

Here, the subscripts t and f refer to the toll road and the free road, and the other notation is the same as in equation (3). The equation means that the second-best toll involves an adjustment to the first-best toll equal to the change in traffic volume on the free road per unit change in traffic volume on the toll road times the marginal congestion cost on the free road. This sign of $dV_f/dV_t$ indicates directly whether the two routes are substitutes or complements. The adjustment to the first-best toll is negative if the routes are substitutes and positive if the two routes are complements.[9] Numerical examples suggested that the optimal toll on the toll road that competes with the free road is about 25% to 45% of the toll that would be charged if both roads were subject to tolls.

## IX. CONCLUSION

This paper has demonstrated that A. C. Pigou made a mistake in his *Wealth and Welfare* in 1912 in asserting that competitive industries subject to increasing costs produce more than the efficient level of output. In his attempt to correct his mistake in the first edition of *The Economics of Welfare* in 1920, he introduced the two-road model and continued to claim that competitive industries subject to diminishing returns to the inputs produce too much. Knight (1924) showed that Pigou's general proposition regarding competitive industries is incorrect, and that the two-road model was not responsive to the question at hand inasmuch as traffic congestion involves an *external* diseconomy. Ellis and Fellner (1943) agreed with Knight, and evidently had the final word on the question of the efficiency of competitive industries. The two-road model was dismissed as not relevant to the primary theoretical issue under discussion. But the transportation economists A. A. Walters (1954), and Beckmann, McGuire, and Winsten (1956) get credit for reviving Pigou's two-road model. Pigou's model (as explained in detail by Knight) became the starting point for the economic analysis of road congestion, road construction, congestion tolls, and related matters.[10]

## REFERENCES

Arnott, Richard, and Kenneth Small. 1994. "The Economics of Traffic Congestion." *American Scientist* 82 (5): 446–455.

Arrow, Kenneth, and Tibor Scitovsky, eds. 1969. *Readings in Welfare Economics*. Homewood, IL:Richard D. Irwin, Inc.

Beckmann, Martin, C. B. McGuire, and Christopher B. Winsten. 1956. *Studies in the Economics of Transportation*. New Haven: Yale University Press.

Cervero, Robert. 2002. "Induced Travel Demand: Research Design, Empirical Evidence, and Normative Policies." *Journal of Planning Literature* 17 (1): 3–20.

Clapham, J. H. 1922. "Of Empty Economic Boxes." *The Economic Journal* XXXII: 305–314.

[9]This result also was derived by Verhoef, Nijkamp, and Rietveld (1996).

[10]The extensive literature on the economics of urban transportation is presented in the textbook by Small and Verhoef (2007). A summary of this field of research is beyond the scope of this paper, which concentrates on the steady-state two-road model.

Downs, Anthony. 1962. "The Law of Peak-Hour Expressway Congestion." *Traffic Quarterly* 16: 393–409.
Duranton, Gilles, and Matthew Turner. 2011. "The Fundamental Law of Road Congestion: Evidence from U.S. Cities." *American Economic Review* 101 (6): 2616–2652.
Ellis, Howard S., and William Fellner. 1943. "External Economies and Diseconomies." *American Economic Review* 33 (3): 493–511.
Knight, F. H. 1924. "Some Fallacies in the Interpretation of Social Cost." *Quarterly Journal of Economics* 38 (3): 582–606.
Leape, Jonathan. 2006. "The London Congestion Charge." *Journal of Economic Perspectives* 20 (2): 157–176.
Marshall, Alfred. [1920] 1961. *Principles of Economics*. Eighth edition. London: Macmillan and Co.
McDonald, John F. 1995. "Urban Highway Congestion: An Analysis of Second-Best Tolls." *Transportation* 22 (4): 353–369.
Pigou, A. C. 1912. *Wealth and Welfare*. London: Macmillan and Co.
Pigou, A. C. 1920. *The Economics of Welfare*. London: Macmillan and Co.
Pigou, A. C. 1922. "Empty Economic Boxes: A Reply." *The Economic Journal XXXII*: 458–465.
Pigou, A. C. 1924. "Comment." *The Economic Journal* 34 (1): 30–31.
Pigou, A. C. [1932] 1961. *The Economics of Welfare*. Fourth edition. London: Macmillan and Co.
Robertson, D. H. 1924. "Those Empty Boxes." *The Economic Journal* XXXIV (1): 16–30.
Robinson, Joan. 1933. *The Economics of Imperfect Competition*. New York: St. Martin's Press.
Robinson, Joan. 1941. "Rising Supply Price." *Economica*, New Series 8 (1): 1–8.
Santos, Georgina, and Erik Verhoef. 2011. "Road Congestion Pricing." in A. dePalma, R. Lindsey, E. Quinet, and R. Vickerman, eds., *A Handbook of Transport Economics*. Northhampton, MA: Edward Elgar, pp. 561–85.
Small, Kenneth, and Erik Verhoef. 2007. *The Economics of Urban Transportation*. New York: Routledge.
Stigler, George J., and Kenneth E. Boulding. 1952. *Readings in Price Theory*. Homewood, IL: Richard D. Irwin, Inc.
Verhoef, Erik, Peter Nijkamp, and Piet Rietveld. 1996. "Second-Best Congestion Pricing: The Case of the Untolled Alternative." *Journal of Urban Economics* 40 (3): 279–302.
Walters, A. A. 1954. "Track Costs and Motor Taxation." *The Journal of Industrial Economics* 2 (2): 135–146.
Walters, A. A. 1961. "The Theory and Measurement of Private and Social Cost of Highway Congestion." *Econometrica* 29 (4): 676–699.
Young, Allyn. 1913. "Pigou's Wealth and Welfare." *Quarterly Journal of Economics* 27 (4): 672–686.

## APPENDIX: SIMPLE MATHEMATICAL MODELS OF RISING SUPPLY PRICE

The first model pertains to a competitive industry that purchases an input with a rising supply price. Assume that the industry uses only one input, labor, which is supplied according to

$$W = w_0 + w'L. \tag{A1}$$

Here W is the wage rate and L is units of labor supplied. The labor supply function is shown in Figure A1.

Further assume that output Q is produced by the industry with constant marginal product:

$$Q = \alpha L. \tag{A2}$$

Total cost for the industry is $TC = WL = WQ/\alpha$, so average cost for the industry is

$$AC = TC/Q = W/\alpha. \tag{A3}$$

The marginal cost of output for the industry is

$$MC = dTC/dQ = W(dL/dQ) + L(dW/dL)(dL/dQ) = (W/\alpha) + (w'L/\alpha). \tag{A4}$$

The total economic rent received by labor is the triangle above the labor supply function, which is triangle $Wbw_0$ in Figure A1. Total rent is, therefore, found to be

$$TR = w'L^2/2, \tag{A5}$$

and the change in total rent as output increases is

$$dTR/dQ = (dTR/dL)(dL/dQ) = w'L/\alpha. \tag{A6}$$

Thus, from equation (A4), we have the result that marginal cost equals average cost plus the change in the economic rent paid to labor.

Equations (A3) and (A4) are two of the cost functions sought. The third cost function is the sum of marginal cost of output and the change in rent, which is

$$MC + dTR/dQ = (W/\alpha) + 2(w'L/\alpha). \tag{A7}$$

Pigou (1912, pp. 172–179) refers to this function as the "difference made to aggregate expenses."

Consider another version of the simple model of industry output produced with one input: labor. Industry output is subject to diminishing returns to the input. The wage is constant $W = w_0$, and output is produced according to

$$Q = L^{\beta}, \text{ with } 0 < \beta < 1. \tag{A8}$$

The marginal product of labor is $\beta/L^{1-\beta}$, and average cost is $TC/Q = w_0L/Q = w_0Q^{(1-\beta)/\beta}$. And marginal cost is

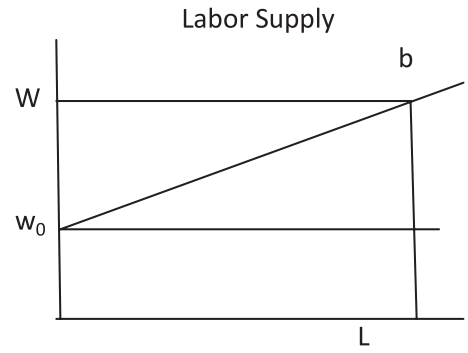


FIGURE A1. Labour Supply

$$MC = dTC/dQ = \left(w_0/\beta\right)Q^{(1-\beta)/\beta} > AC. \tag{A9}$$

The ratio of marginal cost to average cost is simply 1/β, which is greater than one. Marginal cost so computed represents the real additional cost of a unit of output, and is the supply function for the competitive industry. In this case, labor earns no Ricardian rent. Rather, the industry earns Ricardian rent equal to the triangle defined by the equilibrium price p and the marginal cost function, which can be written

$$TR = pQ - w_0 Q^{1/\beta}. \tag{A10}$$

Because P = MC, the change in total Ricardian rent is found as:

$$dTR/dQ = MC + Q\left(dMC/dQ\right) - MC = \left(w_0/\beta\right)\left[\left(1-\beta\right)/\beta\right]Q^{(1-\beta)/\beta}. \tag{A11}$$

Note that:

$$MC - AC = \left(w_0/\beta\right)Q^{(1-\beta)/\beta} - w_0 Q^{(1-\beta)/\beta} = \left(w_0/\beta\right)\left[\left(1-\beta\right)/\beta\right]Q^{(1-\beta)/\beta}. \tag{A12}$$

The change in Ricardian rent equals (MC - AC) × 1/β.

Once again, the change in rent is not a real cost. The real marginal cost is shown in equation (A9). The change in Pigou's aggregate expenses is marginal cost plus the change in Ricardian rent.