"Baumol's Disease": The Pandemic That Never Was

Brian Chapman, PhD BMedSc LMusA
Gippsland Medical School
Faculty of Medicine, Nursing and Health Sciences
Monash University, Australia

Email: Brian.Chapman@med.monash.edu.au
Telephone: +61 3 5122 6225
Facsimile: +61 3 5122 6841

ABSTRACT

Baumol's classic paper entitled "Macroeconomics of Unbalanced Growth: The Anatomy of Urban Crisis" published in the *American Economic Review* in June 1967 is analysed for mathematical and logical errors. Independent errors are found in the derivation of Baumol's Four Propositions and in the analysis of static externalities. It is concluded that (1) none of the Four Propositions can be substantiated because the mathematical derivation lacks sound microeconomic foundations, and (2) the analysis of static externalities falls into algebraic and logical error through failure to link cause with effect and failure to clarify precisely the distinction between intensive and extensive properties.

Key words: Unbalanced growth, static externalities, Baumol's Disease, Baumol effect.

In a much-discussed paper entitled *Macroeconomics of Unbalanced Growth: The Anatomy of Urban Crisis*, William J. Baumol (1967) examined the properties of an economic model in which a technologically progressive sector is differentiated from a stagnant sector. He also included a mathematical analysis of the dependence of static externalities such as pollution and traffic congestion on population growth. The paper drew very pessimistic conclusions about the economic consequences of unbalanced technological progress, predicting ever-increasing costs of education, government services and the fine arts, together with terminal urban blight. The paper attracted widespread interest in the ensuing decades and its bleak scenario acquired epithets ranging from the "Baumol effect" to "Baumol's Disease" and even "Baumol's Curse".[1]

The paper also attracted critical comment from the outset, all directed against the derivation of the Four Propositions and some against the analysis of static externalities. As these two aspects are logically and mathematically independent of each other they will be analysed separately here. The literature critical of the Four Propositions will be discussed after the derivation of the Propositions has been analysed, before moving on to the analysis of static externalities.

I. Baumol's Model of Unbalanced Expansion

A. Preliminary assumptions

Baumol (1967. pp.415-417) stated four initial assumptions as follows:

1. The economy can be divided conceptually into a stagnant sector and a technologically progressive sector (essential).
2. All outlays other than labor costs can be ignored (simplifying, non-essential).
3. Wages in the two sectors go up and down together (essential).
4. Money wage rates rise as rapidly as output per labor unit in the progressive sector (non-essential).

While these assumptions are unobjectionable in themselves,[2] it is noteworthy that Baumol (1967) did not feel it necessary to make or justify here any explicit assumptions concerning either the industrial structure (whether perfectly competitive or otherwise) or the nature of market demand (its shape and variability with time). In principle, it is essential that such assumptions should be made explicit if an internally coherent analysis is to be developed. The following assumptions are therefore added to specify a bare minimum of microeconomic foundations consistent with Baumol's (1967) scenario:

5. the industrial structure is perfectly competitive, and
6. the long-run market demand curve has an explicitly defined shape.

These additional assumptions are consistent with Baumol's (1967) scenario wherein (a) technological benefits are free to expand without limit in the progressive sector and (b) the consequent increased productivity of labour is swiftly rewarded with higher wages in both sectors. Indeed, regarding assumption 5, the historical record suggests that the scenario would be much less likely in non-competitive markets or in a command economy. Unlike some of Baumol's original four assumptions, these two additional assumptions are both implicit in, and essential to, the purported derivation of Baumol's Four Propositions concerning unbalanced growth in a capitalistic economy—indeed they define the essential microeconomic foundations of the argument. Of course, it would be possible, in principle, to assume a non-competitive industrial structure and a highly volatile market demand in the face of technological progress, but this would be contrary to the spirit and intent of Baumol's (1967) derivation. The important point to emphasise is that assumptions of this kind must be made explicit one way or another if the analysis is to have valid microeconomic foundations.

B. Derivation of the Four Propositions

Baumol's (1967) derivation of the Four Propositions commences on p.417 of the original article, thus

> Assume that the economy is divided into two sectors, sector one, in which the productivity of labor is constant, while in sector two output per man hour grows cumulatively at a constant compounded rate, r. Thus we have for the respective values of outputs Y_{1t} and Y_{2t} in the two sectors at time t:
>
> (1) $$Y_{1t} = aL_{1t}$$
>
> (2) $$Y_{2t} = bL_{2t}e^{rt}$$
>
> [p.418] where L_{1t} and L_{2t} are the quantities of labor employed in the two sectors and a and b are constants.

(a) There is an immediate problem here with exactly what is meant by the "respective *values* of outputs". Is it the *number of units* produced or their *monetary value*? Increasing productivity of labour will make possible the output of increased numbers of units for the same input costs. Therefore, equation (2) could possibly express a hypothetical increase in numbers of units produced that might derive from exponentially cumulative technological progress and productivity growth ($r > 0$). However, any cumulative effect is a long-run phenomenon that will necessarily reflect the long-run interactions between supply and demand in competitive markets. This means that the long-run price behaviour and, therefore, the *value* of the increased output will depend on the shape of the market demand curve. Without defining *a priori* what that shape is, there is insufficient information available to accept or reject the next part of Baumol's derivation involving wages.

> We suppose wages are equal in the two sectors and are fixed at W_t dollars per unit of labor, where W_t itself grows in accord with the productivity of sector 2, our "progressive" sector, so that
>
> (3) $W_t = We^{rt}$. (W = some constant)

It is implicit in the argument leading to equation (3) that the expression for the value of output given by equation (2) is intended to describe an exponentially rising real output. Even though Baumol (1967) subsequently clarifies by implication that the symbol for output, *Y*, is expressed in physical units of output (see below), we must assume that the unit price of the outputs remains constant if the real value of output is to rise exponentially and be translated into exponentially rising wages. Thus, we may read equation (2) as purporting to represent either an expression of exponentially increasing *numbers* of physical units produced at constant unit price or, identically, as an expression of exponentially increasing *value* of physical output. Either way, there is no basis for accepting or rejecting the validity of equation (2) unless we know the shape of the long-run market demand curve.

Nonetheless, as Baumol (1967) acknowledges, this growth of wages is not necessary and, in fact, is discarded below where *relative* costs are discussed. However, without this assumption, Baumol (1967) could not have expressed his *Proposition 1* as given (p.418).

> We may now derive several properties of such a system. First and most fundamental is *Proposition 1:* The cost per unit of output of sector 1, C_1, will rise without limit while C_2, the unit cost of sector 2, will remain constant.
>
> Proof:
>
> $$C_1 = W_t L_{1t} / Y_{1t} = We^{rt} L_{1t} / aL_{1t} = We^{rt} / a$$
> $$C_2 = W_t L_{2t} / Y_{2t} = We^{rt} L_{2t} / bL_{2t} e^{rt} = W / b.$$
>
> Note that the *relative* costs will behave in this manner whether or not wages increase in accord with (3) for we have
>
> $$C_1 / C_2 = \left(L_{1t} / Y_{1t}\right) / \left(L_{2t} / Y_{2t}\right) = be^{rt} / a.$$

If wages did not increase, then Baumol's (1967) *Proposition 1* would read, "The cost per unit of output of sector 1, C_1, will remain constant while C_2, the unit cost of sector 2, will decline exponentially."

> In practice, we would expect in these circumstances that market demand for the output of sector 1 would decline (p.418).

Note that this expectation cannot be justified at all on the basis of the above re-expression of *Proposition I* for the case of fixed wages. Indeed, as Robinson's (1969) astute criticism makes clear (see below), the expectation cannot be justified on *either* expression of *Proposition I.*

> Suppose, for example, the elasticity of demand for the two outputs were unity in terms of prices which were proportionate to costs (p.418).

The importance of this supposition cannot be exaggerated. Though, as will be shown, it defines the essential microeconomic foundations of Baumol's scenario, this supposition is not even mentioned, let alone justified, in the introductory section where Baumol (1967, pp.415-417) lists his initial assumptions (see above). Accepting this supposition as given by Baumol (1967) we may therefore rewrite the above assumption (6) more explicitly as

6. market demand remains constant over time and is unit elastic.

This assumption immediately invalidates equation (2) as an expression of *value* because unit elasticity of demand means that the value of output actually demanded remains constant; as more units are produced, the market price falls inversely.[3] The long-run aggregate supply curve[4] for such an industry under these supposed circumstances of constant demand is given by the market demand curve which, by the assumption of unit elasticity, is a rectangular hyperbola. Thus, equation (2) is invalid as an expression of the long-run *value* of that industry's output for all values of $r > 0$. It is also invalid as an expression for the numbers of units of output at fixed unit prices.

> Then relative outlays on the two commodities would remain constant, i.e., we would have
>
> $$\frac{C_1Y_1}{C_2Y_2} = \frac{We^{rt}L_{1t}}{We^{rt}L_{2t}} = \frac{L_{1t}}{L_{2t}} = A \text{ (constant)}$$

The above expression is where Baumol (1967) makes it clear, by implication, that his equations (1) and (2) are expressed in physical units of output.

> Hence the output ratio of the two sectors would be given by
>
> $$Y_1 / Y_2 = aL_{1t} / bL_{2t}e^{rt} = aA / be^{rt}$$
>
> which declines toward zero with the passage of time. Thus we have *Proposition 2:* In the model of unbalanced productivity there is a tendency for the outputs of the "nonprogressive" sector whose demands are not highly inelastic to decline and perhaps, ultimately, to vanish.

One error in this proposition, already noted by others (David F. Bradford, 1969, Michael Keren, 1972, L.K. Lynch and E.L. Redman, 1968) and conceded by Baumol (1972), is that a vanishing output ratio in this case does not mean that the numerator is vanishing but rather that the denominator is increasing exponentially while the numerator remains constant. However, this error becomes irrelevant when we note that this output ratio is derived in terms of equation (2) which has already been invalidated by the assumption of unit elasticity. In short, the output ratios are unchanged where output is properly defined in terms of real value.

> We may inquire, however, what would happen if despite the change in their relative costs and prices the magnitude of the relative outputs of the two sectors were maintained, perhaps with the aid of government subsidy, or if demand for the product in question were sufficiently price inelastic or income inelastic. Then we would have
>
> $$b / a = Y_1 / Y_2 = L_1 / L_2e^{rt} = K.$$

Not only do the *relative* outlays remain constant under the assumption of unit elasticity, so also do the *absolute* outlays if wages do not rise; and even if wages do rise then the *real* outlays remain constant. Moreover, the ratio of the outlays will remain constant whether measured in either nominal or real terms.

> Let $L = L_1 + L_2$ be the total labor supply. It follows that
>
> (4) $$L_1 = (L - L_1)Ke^{rt} \quad \text{or} \quad L_1 = LKe^{rt}/(1+Ke^{rt})$$
>
> and
>
> (5) $$L_2 = L - L_1 = L/(1+Ke^{rt})$$
>
> Hence, as *t* approaches infinity, L_1 will approach L and L_2 will approach *zero*. Thus we have *Proposition 3:* In the unbalanced productivity model, if the ratio of the outputs of the two sectors is held constant, more and more of the total labor force must be transferred to the non-progressive sector and the amount of labor in the other sector will tend to approach zero.

This does not follow. Given unit elasticity of demand in both sectors, if the ratio of the outputs of the two sectors is held constant (in real value), more and more *units* of sector 2 output will be consumed at constant aggregate value of consumption while consumption of sector 1 output will remain constant both in value and number of output units consumed. The proportions of labour employed in the two sectors will remain constant.

> Finally, we may note what happens to the overall rate of growth of output in the economy if the output ratio for the two sectors is not permitted to change. We may take as an index of output a weighted average of the outputs of the two sectors:
>
> $$I = B_1Y_1 + B_2Y_2 = B_1aL_1 + B_2bL_2e^{rt}$$
>
> so that by (4) and (5)
>
> $$I = L(KB_1a + B_2b)e^{rt}/(1+Ke^{rt}) = Re^{rt}/(1+Ke^{rt})$$
>
> where
>
> $$R = L(KB_1a + B_2b)$$
>
> Therefore
>
> $$dI/dt = R\left[re^{rt}(1+Ke^{rt}) - Ke^{2rt}\right]/(1+Ke^{rt})^2$$
> $$= rRe^{rt}/(1+Ke^{rt})^2$$
>
> As a result, the percentage rate of growth of output will be
>
> $$\left(dI/dt\right)/I = r/(1+Ke^{rt})$$
>
> which declines asymptotically toward zero as *t* increases. We have, then, arrived at *Proposition 4:* An attempt to achieve balanced growth in a world of unbalanced productivity must lead to a declining rate of growth relative to the rate of growth of the labor force. In particular, if productivity in one sector and the total labor force remain constant the growth rate of the economy will asymptotically approach zero.

Under the unit elasticity assumption, there can be no unbalanced growth of one sector relative to the other regarding the long-run proportions of GDP. Nor is there any absolute growth in the *real value* of GDP. Only the *quantity* demanded from the progressive sector rises, not the *value* of the quantity demanded. Baumol's (1967) derivation collapses because it is contradicted by his supposed microeconomic foundations. The assumption of unit elastic demand invalidates equation (2) and every equation and proposition derived from it.

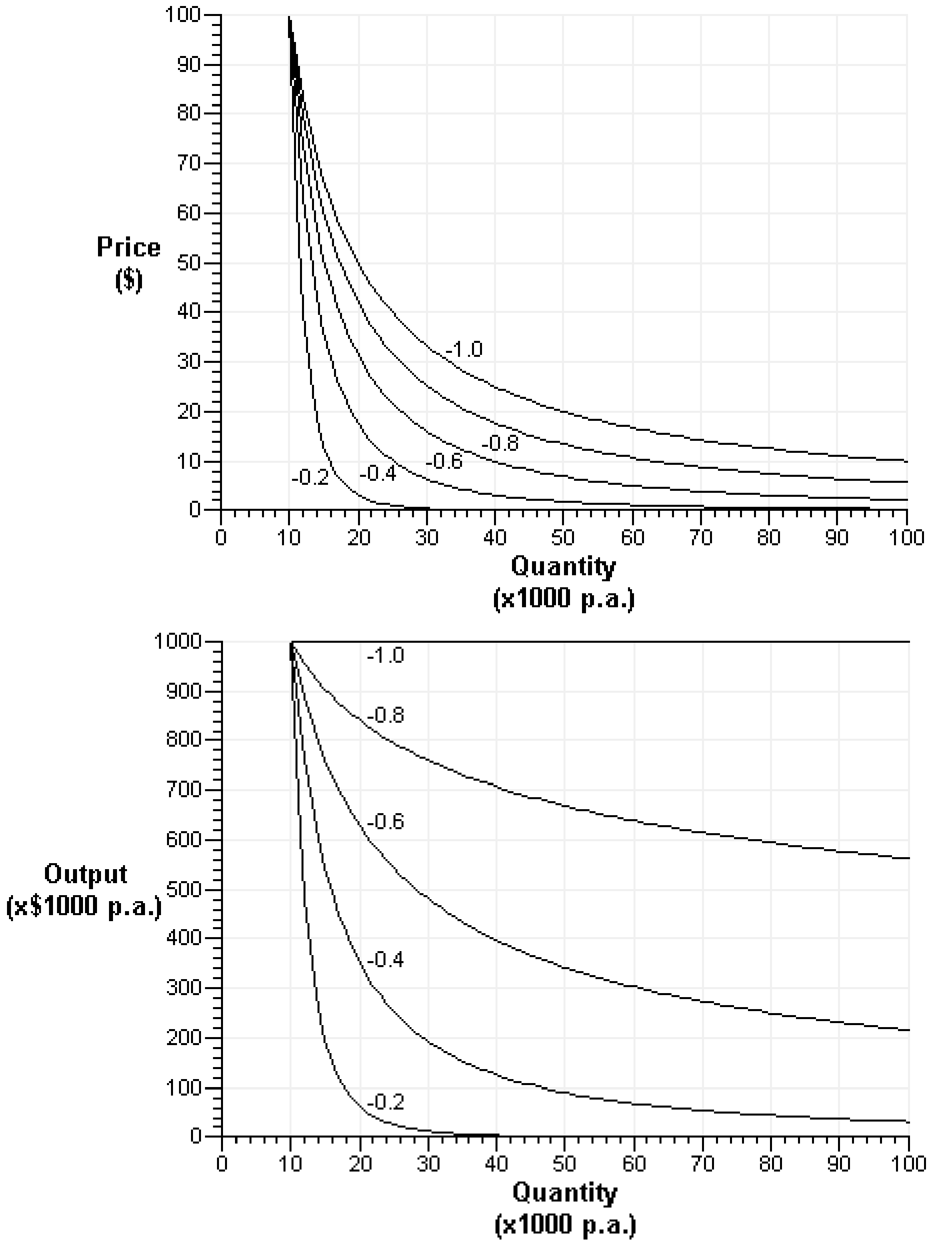

Figure 1. Long-run demand (upper) and output (lower) for different price elasticities

C. Discussion

The assumption of constant unit elasticity of demand is not particularly compelling; if the watch industry were taken as an exemplar for the progressive sector (William J. Baumol et al., 1985, p.87), then we have difficulty imagining a situation in which the number of watches consumed *per capita* would increase ‘without limit’ as the price of watches fell exponentially.[5] Given that real-world demand for technologically progressive goods is generally less than unit elastic, a technologically progressive industry will necessarily produce less *value* of output per worker as time goes by unless it sheds labor.

This is illustrated in Figure 1 which shows hypothetical demand curves (upper) and output curves (lower) for five different prices elasticities ranging from –0.2 to –1.0. The long-run scenario for each curve is assumed to commence at a unit price of $100 and an equilibrium quantity demand of 10,000 units *per annum* and to be followed by increased productivity of labor with fixed labor inputs. Thus, if technological progress is to be viewed with pessimistic suspicion at all, then our concerns should be with rising unemployment and *increasing demands upon the workers remaining in employment in the progressive sector* rather than on the safe, if unexciting, employment prospects continuing in the stagnant sector.

However, the assumption of any kind of *constant* elasticity of demand, sustained over cumulative generations of technological progress, is difficult to justify.[6] But without an explicit assumption about demand elasticity Baumol's purported macroeconomic scenario cannot be analysed properly.

D. Conclusion

The derivation of the Four Propositions is thus an unfortunate example of the misapplication of mathematics to an inadequately specified and unrealistically simplified economic problem (see J. Blatt, 1983). None of the Four Propositions for the behaviour of a model of an unbalanced economy set forth by Baumol (1967) can be supported. Their mathematical derivation is invalidated at the outset by any sort of tractable assumptions for the necessary microeconomic foundations of the model. These foundations received insufficient attention and explication in Baumol's (1967) original exposition.

E. The Critics of the Four Propositions

This section deals with the criticisms contained in several 'comments' by various authors published in the *American Economic Review* during the period 1968-1972 and in a separate paper published in *Zeitschrift für Nationalökonomie* in 1969, together with three responses published by Baumol in the *American Economic Review* over the same period.

1. Bell, 1968

Bell (1968) drew attention to the role played by the "technology of consumption" (p.877) and noted that consumers' tastes "stand in the same relation to any given demand function as does technology to any given supply function" (p.879). Regarding the productivity of services, Bell (1968) also quoted examples of services that come into existence "because the technologies of using goods, as well as manufacturing them, have changed" (p.883). The ultimate futility of comparing productivity in the goods and services sectors was recognised in Bell's (1968) assertion that "the only proper way to evaluate any industry is to set up an input-output ratio consisting of labor and consumption utility" (p.883). Thus, it was concluded (Carolyn Shaw Bell, 1968, p.884):

> If it were possible to quantify consumer satisfaction, the output of both goods and services could be measured in comparable terms and then productivity, defined as the ratio of labor input to utility, could be compared among industries or sectors of the economy. The absence of such data, however, indicates that great caution must be exercised in describing the relative growth of goods-producing industries and the service sector.

Bell's (1968) objections therefore pertain to

(a) the fact that technological progress in relation to consumption is just as important a component of economic progress as it is in relation to production, and

(b) the difficulties of comparing productivity between the goods sector and the services sector

without touching upon the problem of defining appropriate microeconomic foundations for Baumol's (1967) scenario.

2. Lynch and Redman, 1968

Lynch and Redman (1968) focused their criticism on their claim that Baumol "overlooked the fact that *real income* … is rising", leading them to suggest (p.885)

> Only those nonprogressive goods for which the absolute value of the elasticity of demand with respect to price is less than the absolute value of income elasticity, or those for which the income effect is negative, will tend to vanish.

They then noted that declining outputs of costly service industries (performing and fine arts, gourmet dining)—which should all show high income elasiticity—might be due in part to changing distribution of income and changing tastes as much as to relative prices changes.

However, by failing to take into account the necessity of defining properly the microeconomic foundations of the problem, Lynch and Redman (1968) left themselves with no theoretical or practical justification for their assertion that real income is rising. Quite simply, their can be no increase in either nominal or real income if the unit price of output falls inversely as the number of units produced per worker increases. Nonetheless, if a properly defined scenario involving rising real incomes were to be conceived, then Lynch and Redman's (1968) criticisms would become relevant.

3. Worcester, 1968

Worcester (1968, p.887) had "no quarrel with" Baumol's (1967) "simplifying assumptions nor with the logical steps" but rather with the claimed implications. He constructed a diagram illustrating various growth paths, production possibilities and indifference curves, all serving to convey a less bleak set of options than was given in Baumol's (1967) original scenario. Worcester (1968, p.890) noted that, as production capacity for the progressive good (y) grows, its price falls "again and again." Consequently, he continued, "Anxious eyes now scan the horizons for adequate markets for the flood of y that rises progressively and cumulatively." Nonetheless, he did not make the crucial connection between this market behaviour and the definition of the microeconomic conditions essential to a proper analysis of the macroeconomic problem.

4. Birch and Cramer, 1968

Birch and Cramer (1968) purported to show "that unit costs in the non-productive sector may remain constant or even decline under an assumption of wage diffusion when any reasonable demand constraint is allowed" (p.893). The explicit demand constraint they considered was that the ratio of total spending in the two sectors would be equal to the sectoral ratio of total costs (p.894) and that this ratio would be constant over time (p.895, equation 12). They argued that the strength of a wage diffusion effect "has to depend on the relative proportion of labor in the productive sector" (p.894). While these authors cited empirical support for a "decreasing relative output from the productive sector" (p.896), allowing a case to be made "for a decline in the unit cost of services" (p.895), they did not express their definitions of output rigorously enough and they overlooked the internal contradictions of Baumol's (1967) analysis.

5. Baumol, 1968

Baumol (1968) did not contest directly any of the explicit criticisms published in the above four papers but rather took "the opportunity to clear up a few misunderstandings" which, he claimed, "characterized a number of comments on" his original paper. His unshaken commitment to his original scenario was confirmed thus (William J. Baumol, 1968, p.897):

> I persist in the view that the extraordinary rates of cost increase that we observe throughout the nonprogressive sector is a matter of their technology—the special role of the labor input in determining the quality of their end product.

6. Robinson, 1969

The 'belated' comment published by Joan Robinson (1969, p.632), despite its brevity (half a page), was the only early criticism that went straight to the heart of the problem as identified in the present work. With understated eloquence she explained:

> As output per man employed in industry rises, the prices of commodities sold to the public will fall, while the price of services remains constant. The cost of services in terms of commodities is rising just because the cost of commodities in terms of labor is falling.

There is nothing in the present analysis that goes beyond, or is not implicit or latent in, the intuitive understanding conveyed in this simple statement. Nonetheless, Robinson's (1969) statement did not serve to prevent the spread of Baumol's (1967) misconceived scenario in the macroeconomics literature.

7. Baumol, 1969

Baumol's (1969, p.632) half-page response to Robinson's (1969) comment is unsatisfactory and its essential content should be reproduced in full, thus:

> If we determine the real cost of services by the amount of labor needed to produce them, that cost is clearly not increased by the declining labor cost of manufactures. But, if we measure the real cost of a unit of service by its *opportunity cost*, the quantity of manufactures one must give up to obtain it, my proposition holds. It is obviously the latter definition I had in mind, though I should, no doubt, have said so more clearly.

On the contrary, it is obvious that Baumol (1969) has not answered Robinson's (1969) criticism at all but has merely restated it while apparently betraying an unfortunate misunderstanding of what economists mean by the concept of "opportunity cost". If labour productivity doubles, then the number of widgets one must give up to purchase a haircut may well double, but the amount of labor one must expend in order to produce the double quantity of widgets does not increase at all—hence, there is no increase in real opportunity cost.

8. Bradford, 1969

Bradford (1969) suggested that Baumol's (1967) rather pessimistic Four Propositions could be "balanced by rather more positive ones" (p.304), thus:

> Proposition 1 (Baumol): Through time the cost per unit of output of the lagging sector 1, relative to that of the progressive sector 2, will rise without limit.
>
> Proposition 1' (Bradford): Through time the cost per unit of output of sector 2, relative to that of sector 1, will fall to zero in the limit, while the set of feasible output combinations will continually expand.

This restatement is appropriate and emphasises the direct relation of cause and effect, i.e., the *effect* on costs in the progressive sector *caused* by the technological progress. Bradford (1969) also deals with expansion paths and production possibilities after the manner of Worcester (1968).

> Proposition 2 (Baumol): Unless demand for it is highly inelastic, output in the lagging sector will decline and perhaps vanish.
>
> Proposition 2' (Bradford): By maintaining a constant proportion between expenditures on the two goods (at current prices), individuals can purchase a constant amount of Y_1 and an ever growing amount of Y_2.

This is another appropriate restatement that identifies causes and effects appropriately in the changing (progressive) sector. Bradford (1969) also touches upon difficulties with what he sees as an implicitly defined "hybrid elasticity" (as between income- and price-elasticity) in Baumol's (1967) analysis without identifying the fundamental problem associated with the assumption of unit elasticity.

> Proposition 3 (Baumol): In order to maintain balanced expansion (i.e., a constant ratio between the outputs of the two sectors), the fraction of the labor force allocated to the lagging sector must approach unity.
>
> Proposition 3' (Bradford): It seems unlikely that individuals (or society collectively) will choose a balanced expansion path in a world of unbalanced productivity change.

This restatement is reasonable and underlines the importance of explicit assumptions about market demand as defining the microeconomic foundations of the problem, without which foundations no macroeconomic conclusions can be drawn.

> Proposition 4 (Baumol): An attempt to maintain balanced expansion must lead ultimately to a zero rate of growth of real output per man.
>
> Proposition 4' (Bradford): As long as any ceiling, however close to 1 (but less than 1), is placed on the fraction of the labor force allocated to the lagging sector, the average rate of growth of real output per man over the long run will equal the maximum sustainable level, which is the rate of growth of productivity in sector 2.

This restatement does not address the fact that the unit elasticity assumption means that there is no growth in real productivity in sector 2. Bradford (1969) also criticised the weakness and artificial limitations resulting from Baumol's (1967) assumption of only one factor as an input for production. However, these criticisms stopped short of identifying the fundamental logical inconsistency of the supposed unit elasticity of demand in Baumol's (1967) analysis.

9. Keren, 1972

Keren (1972) directly attacked the falsehood of Baumol's (1967) Proposition 2 as stated, correctly asserting that "the output ratio of" the stagnant sector to the progressive sector "falls because the denominator is rising, not because the numerator is falling." He also claimed that "several of Baumol's conclusions depend on the false Proposition 2" (p.149), giving education and municipal services as examples. Concerning the former he wrote (p.149):

> It does not follow that the relative outlays on higher education should rise, unless relatively more or better education is being provided.

Concerning the latter he wrote (p.149):

> … only if the income elasticity of municipal taxes is below unity will the financial problems of cities grow, provided the level of services remains unchanged. This could happen if the municipal tax base does not grow as fast as labor income.

It remains only to be added that the ubiquitous progression that exists in nearly all democratic taxation structures firmly rules out any such contingency for urban financial problems.

10. Baumol, 1972

In a half-page reply to Keren's (1972) note, Baumol (1972) acknowledged the correctness of the criticism and drew attention to the similar criticisms raised by Lynch and Redman (1968) and Bradford (1969). It is significant that he did not take the opportunity to confirm that the much more penetrating criticism raised by Robinson (1969) really subsumes all these niceties and makes them essentially redundant.

Nonetheless, in seeking to defend "Baumol's Disease", Baumol (1972) continued (p.150):

> The basic argument of the original analysis still remains valid: the financial problem of the cities increases (in part) because the costs of the services rise more rapidly than the general price level. This may well lead to cumulative deterioration in the quality and quantity of the services, even though it is not forced on the economy by lack of resources.
>
> Because of their rising relative costs, the community may in fact *want* less of some or even most of these services. But the danger is that the nature of the political process and the tax system will also force it to accept a decline in even those services which the public would in fact prefer to expand, an expansion which, as Keren shows, society will well be able to afford.

No support is offered for this statement which is, in fact, insupportable in the face of the progressive taxation structures that determine the public revenues of democracies. Any tendency for technological progress to increase national income (whether in nominal or in real terms) will automatically channel disproportionately more resources to the public sector through progressive taxation. The problem is not one of inadequate financial resources for cities, but rather one of ensuring that the unlegislated growth of public revenue is applied wisely where it is needed or refunded to the taxpayer.

11. Summary

As a matter of historical record, it has to be admitted that none of the above criticisms has been effective in causing the profession of macroeconomics to disregard "Baumol's Disease" as an important influence in the economic growth of advanced societies. Not one of the criticisms explicitly identified the fundamental lack of sound microeconomic foundations in Baumol's (1967) scenario or the logical and mathematical inconsistency between the casual supposition of unit elasticity of demand and the exponentially increasing value of output in the progressive sector. We have already noted that Robinson's (1969) criticism did this implicitly, but the implication was too subtle to have the deserved effect.

Baumol subsequently revisited his scenario and added to it the concept of 'asymptotically stagnant activities'—activities that initially comprise a significant progressive component alongside a stagnant component, the latter coming to dominate over time as costs and prices fall in the progressive component (William J. Baumol, Sue Anne Batey Blackman and Edward N. Wolff, 1985). This paper made neither reference nor concession to any of the earlier criticisms, continuing with the unswerving conclusion that:

> In an economy in which the productivity growth rates of the different sectors are unequal, it is impossible for both the output ratios and the input ratios to remain constant (p.807).

This conclusion cannot be accepted without knowing what exactly is meant by 'output' and 'productivity' and in what units they are expressed. In short, all the criticisms raised in the present paper remained unanswered, along with the earlier criticisms, especially that of Robinson (1969).

II. Static Externalities

Baumol (1967, 423-424) set out what he claimed to be a 'natural premise' concerning the growth of externality costs:

> Now there are undoubtedly many reasons for the explosion in external costs but there is a pertinent observation about the relationship between population size in a given area and the cost of externalities that seems not to be obvious. It is easy to assume that these costs will rise roughly in proportion with population but I shall argue now that a much more natural premise is that they will rise more rapidly—perhaps roughly as the square of the number of inhabitants.

He then proceeded to give two examples pertaining to environmental pollution and traffic congestion, respectively.

A. Air pollution

> For example, consider the amount of dirt that falls into the house of a typical urban resident as a result of air pollution, and suppose that this is equal to *kn* where *n* is the number of residents in the area. Since the number of homes in the area, *an*, is also roughly proportionate to population size, total domestic sootfall will be equal to soot per home times number of homes = $kn \cdot an = akn^2$ (Baumol, 1967, p.424).

The error in this argument is that there has been an unwarranted 'double-counting' of the influence of population density, with two correlated quantities—soot/population and homes/population—being treated as mathematically independent variables. This is quite wrong because the homes are the *source* of the soot in the first place. Such 'double-counting' errors are avoided by properly accounting for the physical relations between cause and effect. We must also address separately the two aspects of population density—the density of housing and the size of the household.

1. Housing density

Consider an idealised housing development in which every home is identical and occupied by a single person who uses a log fire for domestic heating, cooking and bathing, generating *S* amount of soot annually. Let us suppose that the soot from this fire ascends and spreads in such a way that all of it falls within the property boundary of each household, each property being of just sufficient size to ensure that there is no spill of soot from any property to any of its neighbours as represented in Figure 2 (a).[7] Thus, the annual soot generated by each household is *S* while the annual sootfall experienced by each household is also exactly *S* because each household receives 100 percent of its own soot and no soot from its neighbours.

Now suppose that the housing density is increased slightly to that represented in Figure 2 (b). The soot from each household will now fall partially on its own acreage and partially spill over onto neighbouring properties, while each household will also receive soot precipitation from its neighbours. No matter how dense the housing development becomes, as represented in Figure 2 (c) and (d), the total sootfall experienced by each household can never exceed the soot generated per household; the amount of soot 'lost' by each household onto neighbouring properties will be exactly balanced by incoming soot spilling over from those same properties.[8]

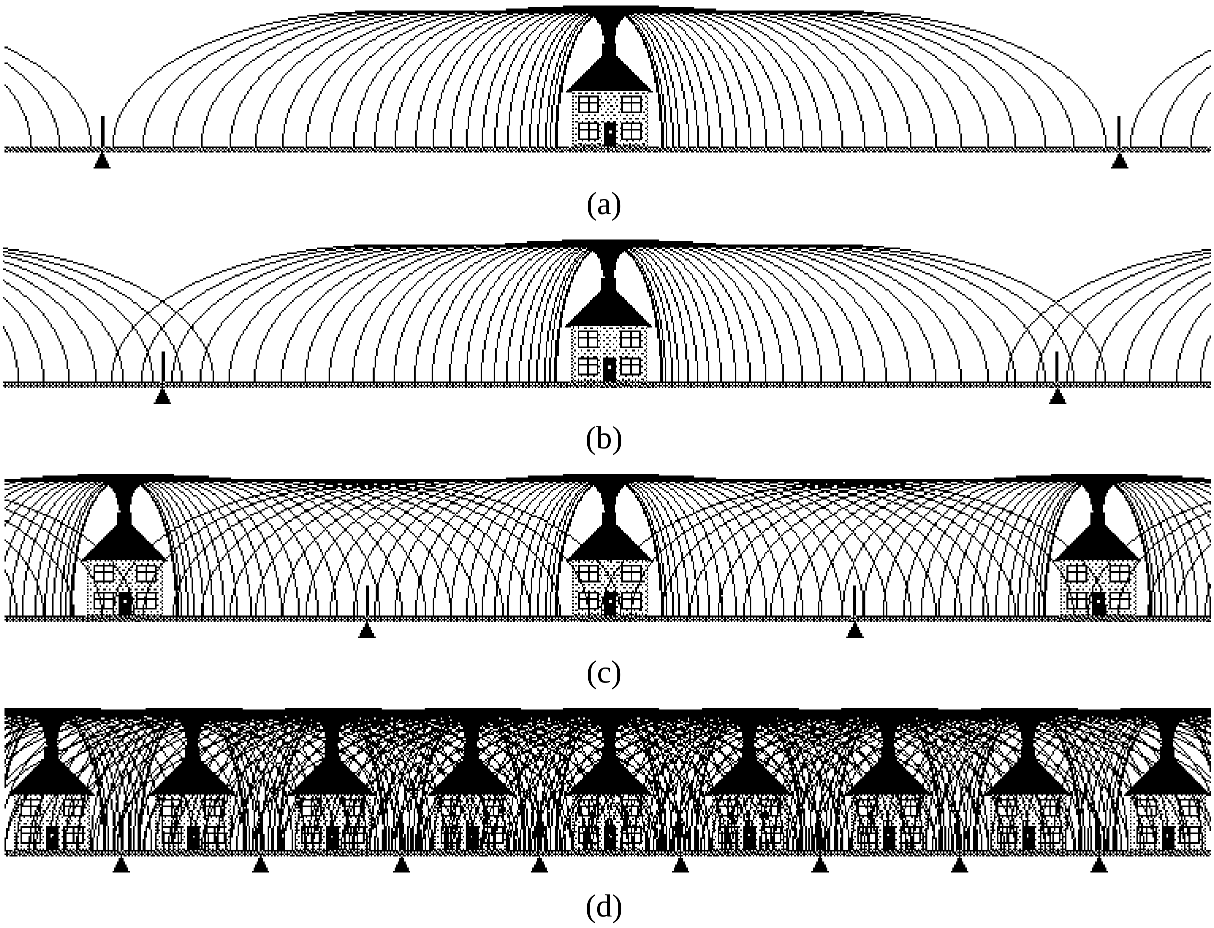

Figure 2. Dependence of soot generated and sootfall received on housing density

Thus, ignoring winds and air currents, we may safely conclude for our idealised model of identical households located on identically-sized blocks of land and housing identically-behaved single occupants, that the average sootfall experienced *per house block* at the centre of a large urban development is equal to the average soot generated per household and is independent of the size of the identical blocks. This implies that the average sootfall per house block at the centre of town is *independent of housing density*, *i.e.*, it is *zero-order* with respect to housing density. On the other hand, the sootfall experienced per house block at the edge of town *decreases* with increasing density of housing as the decreasing size of each house block results in more spillage onto the surrounding countryside.

2. Household size

If we allow the number of persons per household to increase then we should expect a *less than proportionate* increase in the amount of soot generated per household. This is because the amount of soot generated *per person* by household heating, cooking, and bathing is a decreasing function of the number of persons per household.

Once a house has been heated sufficiently for the comfort of a single occupant, the marginal fuel consumption required for the comfort of additional occupants is much less than the fuel consumption required by a single occupant. Similar considerations apply to the provision of heat for cooking and bathing, though the relative marginal fuel consumed per person might be expected to decrease in the order bathing > cooking > heating.

3. Population density

From the foregoing arguments we may conclude that the sootfall experienced per household

1. is independent of housing density at the centre of town and a decreasing function of housing density at the edge of town,
2. has a less than first-order dependence on the number of persons per household, and
3. therefore, has overall a less than first-order dependence on population density.

4. Aggregate pollution costs and per capita expenditure

Let us suppose that our idealised community wished to control its soot emissions by installing afterburners or some other pollution control devices on all domestic chimneys. It is reasonable to expect that the costs of renting and operating such equipment would show similar dependence on population density as for the original soot generation process, with constant fixed costs per household to rent and operate a unit for single-occupancy dwellings, and a less than proportionate increase in these costs for larger capacity units appropriate to multiple-occupancy dwellings.

The *per capita* expenditure on pollution control would thus be less than first order with respect to population density, well short of the second-order relationship suggested by Baumol (1967). Even the aggregate community expenditure, an *extensive* quantity, would be less than first-order for this example (see Table 1). However, it is highly misleading to focus attention on extensive quantities. Any externalities problem is experienced by individuals and paid for by individuals on a *per capita* basis; it is the *per capita* measurement of the problem and its cost which is relevant to the problems associated with increasing urban size.

5. Air pollution and its costs of elimination

While we concluded above that the sootfall experienced per house at the centre of our large idealised urban development is independent of the housing density, this conclusion does not apply to the quality of the air being breathed by the citizens. The more tightly packed are the houses, the more soot is being borne per unit volume of air around the houses as the constant amount of soot reaching any given house is concentrated into house blocks having smaller area (see Figure 2). This would suggest something approaching a first-order dependency between air pollution and housing density, with a less than first-order dependency between air pollution and household size (see Table 1).

The overall relation between air pollution and population density would depend on the exact way in which the population density increased. If the increase were due solely to an increase in housing density with no increase in household size, then the relation would be first-order. On the other hand, if the increase were due solely to increasing household size with no increase in housing density, then the relation would be less than first-order. Thus, given a mixture of both increasing household density and increasing household size, the overall relation between air pollution and population density should be somewhat less than first-order. Nonetheless, the relation to the population density of the *per capita cost* of minimising or eliminating this pollution is as already given in the preceding section; it is identical with the relation to the population density of the *per capita generation* of the pollution and is less than first-order. If this cost were borne by the municipal authorities, the required relative growth of the relevant division of the public sector would be less than first-order in its dependence on population size.

B. Traffic congestion

> Similarly, if delays on a crowded road are roughly proportionate to *n*, the number of vehicles traversing it, the total number of man hours lost thereby will increase roughly as n^2, since the number of passengers also grows roughly as the number of cars" (Baumol, 1967, p.424).

It is reasonable to suppose that the average duration of a delay for a vehicle on a given highway is positively correlated with the number of vehicles in transit per unit length of road. Whether the correlation is first-order or higher-order is questionable, but we shall accept Baumol's (1967) suggestion that it is roughly first-order. However, we have greater difficulty with the ambiguous statement that the "number of passengers also grows roughly as the number of cars". Taking the designation 'passengers'[9] to include the drivers of the vehicles, this statement could possibly mean either of the following:

(a) The number of persons *per vehicle in transit* grows roughly as the number of cars in transit per unit length of road;

(b) The number of persons *in transit per unit length of road* grows roughly as the number of cars in transit per unit length of road.

Alternative (a) seems unlikely. The only condition that would simultaneously increase the number of cars in transit *and* the number of persons per vehicle in transit would be a large shift of commuters from public transport to private *car-pooling*. Any shift of commuters from private individual transport to private car-pooling would *decrease* the number of vehicles in transit while raising the number of persons per vehicle in transit. Increases in the number of vehicles in transit in proportion to increases in the population will not increase the number of persons per vehicle in transit, while increases in the number of vehicles per head of population will tend to reduce the average number of persons per vehicle in transit.

On the other hand, alternative (b) is trivial, given that each car in transit requires a driver who is included in the classification 'persons in transit' for the purposes of accounting for the number of man-hours lost through traffic congestion. Nonetheless, let us assume that the number of persons travelling per unit length of road, *n* is proportionate to *C*, the number of cars in transit per unit length of road, such that

(6) $$n = k_1 C$$

and the average duration of delay, *d* is also proportionate to *C*, such that

(7) $$d = k_2 C$$

As the average delay for each vehicle is thus $k_2 C$ hours and the average number of persons delayed per vehicle is k_1 persons, the average number of man-hours lost per vehicle, *D*, is given by

(8) $$D = k_1 d = k_1 k_2 C$$

Equation (8) defines a first-order[10] relation between man-hours of delay per vehicle and traffic density. This is an *intensive* measure of the consequences of traffic density. The *extensive* measure of aggregate loss of man-hours for the entire population will, of course, be second-order with respect to traffic density, but this immediately reduces to a first-order dependence as soon as *per capita* costs of untreated traffic congestion are estimated. On the other hand, the costs of *avoiding* traffic congestion could be imposed as fixed charges on vehicle usage (e.g., tolls), implying a *zero-order per capita* cost of traffic density management (see Table 1).

Thus Baumol's (1967) finding of a second-order relationship between the number of man-hours lost through delays and the density of vehicles in transit applies only to the *extensive* aggregate quantity under consideration, *i.e.*, the aggregate number of man-hours lost for the entire population. It does not lend support to the notion that the *per capita* cost of dealing with the problem is second-order with respect to traffic density or population density.

C. Generalisation

> The logic of the argument is simple and perhaps rather general: if each inhabitant in an area imposes external costs on every other, and if the magnitude of the costs borne by each individual is roughly proportionate to population size (density) then since these costs are borne by each of the *n* persons involved, the total external costs will vary not with *n* but with n^2 (William J. Baumol, 1967, p.424).

This sentence confuses intensive and extensive properties, implying erroneously an equivalence between 'population size' and 'population (density)'. It may be clarified in this respect if re-phrased as: "If each inhabitant in an area imposes external costs on every other, and if the magnitude of those costs *per capita* is roughly proportionate to population density, *n*, then the *aggregate* external costs will vary with n^2." This corresponds roughly to the above example of traffic density, but it does not relate to the above example of air pollution.

However, the unfortunate use of extensive quantities (size) becomes rather confusing as Baumol concludes his generalisation.

> Of course I do not maintain that such a relationship is universal or even that it is ever satisfied more than approximately. Rather I am suggesting that, typically, increases in population size may plausibly be expected to produce disproportionate increases in external costs—thus pressures on the municipality to do something about these costs may then grow correspondingly (Baumol, 1967, p.424).

We reject entirely, in terms of the examples given, the notion that 'increases in population *size*' will produce disproportionate increases in *per capita* external costs. If it be contended that Baumol (1967) was referring to *aggregate* external costs—an *extensive* quantity, then we suggest that this is a misleading way in which to describe external costs; such a description runs the risk of creating opportunities for political mischief rather than rational policy-making.

D. Conclusion

We are driven to reject Baumol's (1967) "natural premise" that static urban externalities will rise "perhaps roughly as the square of the" population density. The examples given do not show second-order relations for air pollution on population density or traffic delays on traffic density, particularly when *per capita* externalities and abatement costs are considered. The errors here are different from those pertaining to the derivation of the Four Principles. In the case of the Four Principles, the mathematics was flawless but unfounded; in the case of the static externalities both the logic and mathematics were inherently erroneous.

E. The Critics of the analysis of static externalities

1. Lynch and Redman, 1968

Although Lynch and Redman (1968, p.885) did not identify explicitly the same logical and mathematical errors in Baumol's (1967) analysis as we have done here, they drew attention nonetheless to the importance of analysing the costs of municipal services on a *per capita* basis.

2. Worcester, 1968

Worcester (1968, p.892) found no fault with “Baumol’s analysis of air pollution and traffic problems which finds costs rising by the square of the population density” but rather expressed disagreement with the implications for policy. For example, he questioned “the purpose of subsidized rapid transit” that “is bound to increase the population density near its terminals.” He also noted that externalities problems are more appropriately addressed by focusing on the costs to individuals of implementing the various remedies, and that urban diseconomies are frequently offset by accompanying economies in the delivery of health, education and welfare services.

III. Discussion and Overall Conclusion

The reader is invited to reflect again on the passage quoted from Robinson (1969) on page 9. Had such a concise understanding of the classical law of markets been held by macroeconomists as an unshakeable axiom in the 1960s, it is unlikely that the idea of "Baumol's disease" would have found a locus in which to germinate, let alone an elite journal from which to influence the profession. But macroeconomists in the late 1960s were still largely in thrall to the 'anti-classical' Keynesian *zeitgeist*, and even the monetarists (the purported 'anti-Keynesians') lost the initiative as their chief apostle proclaimed that they were "all Keynesians now" (Milton Friedman, 1968, p.15). So it was that Robinson's (1969) antidote failed to stop the 'disease' in its tracks.

Today, "Baumol's Disease" is entrenched in the macroeconomics literature as a significant factor, among others, 'explaining' the disproportionate growth of government (dominated by 'services') in the latter half of the twentieth-century (e.g., see Johan A. Lybeck, 1986, Chapter 5, Dennis C. Mueller, 1989, Chapter 17). Thus, given the logical and mathematical errors embedded in the concept and derivation of "Baumol's Disease", it stands as a monument to perpetuation of error in macroeconomic analysis sustained over more than forty years.

What, then, are we to make of studies that claim to find empirical evidence for a "Baumol effect" in the rising cost of services generally and government services in particular?[11] The answer is logically straightforward: it is no vindication of a flawed theory if its predictions happen to accord with empirical results; rather, the problem is to remedy the deficiencies in the original theory or to find alternative, valid explanations for the results. Fortunately, there is no shortage of demand-side and supply-side explanations available for the growth of government (e.g., see Johan A. Lybeck, 1986, Chapter 5, Dennis C. Mueller, 1989, Chapter 17). These explanations need no further elaboration here beyond the suggestion that insufficient recognition has been given to an important supply-side mechanism—the unlegislated growth of progressive taxation that persisted for decades throughout the inflationary period in which "Baumol's Disease" was conceived, published, discussed and studied. The potential for this supply-side effect on the growth of government has been long known to economists and was authoritatively confirmed in an ominously prophetic observation more than sixty years ago:

> For inflation is a mighty tax-gatherer (John Maynard Keynes, 1940, p.68).

Acknowledgment

The author wishes to acknowledge helpful discussion and criticism of earlier drafts of this paper provided by Professor Paresh Narayan and an anonymous reviewer.

TABLE 1 — Dependences of intensive and extensive properties of externalities on population

AIR POLLUTION

Intensive Externality or Cost	*Dependence on*		
	Housing density	*Household size*	*Population density*
Soot received per household	Zero-order	Less than first-order	Less than first-order
Soot per unit volume of air	First-order	Less than first-order	Less than first-order
Average abatement cost per household	Zero-order	Less than first-order	Less than first-order

Extensive Externality or Cost	*Dependence on*		
	Total number of houses	*Household size*	*Total population*
Total soot *received* by all households	First-order	Less than first-order	Less than first-order
Aggregate abatement costs	First-order	Less than first-order	Less than first-order

TRAFFIC CONGESTION

Intensive Externality or Cost	*Dependence on*		
	Traffic density	*Persons per vehicle*	*Travelling population density*
Man-hours lost per vehicle	First-order	First-order	First-order
Road tolls per vehicle	Zero-order	Zero-order	Zero-order

Extensive Externality or Cost	*Dependence on*		
	Total number of vehicles	*Persons per vehicle*	*Total travelling population*
Aggregate man-hours lost	Second-order	First-order	Less than second-order
Aggregate toll charges	First-order	Zero-order	Less than first-order

References

Baumol, William J. "Comment on the Comment." *American Economic Review*, 1969, *59*(4), pp. 632.
____. "Macroceonomics of Unbalanced Growth: Comment." *American Economic Review*, 1968, *58*(4), pp. 896-97.
____. "Macroeconomics of Unbalanced Growth: The Anatomy of Urban Crisis." *American Economic Review*, 1967, *57*(June), pp. 415-26.
____. "Macroeconomics of Unbalanced Growth: Reply." *American Economic Review*, 1972, *62*(1/2), pp. 150.
Baumol, William J.; Blackman, Sue Anne Batey and Wolff, Edward N. "Unbalanced Growth Revisited: Asymptotic Stagnancy and New Evidence." *American Economic Review*, 1985, *75*(4), pp. 806-17.
Bell, Carolyn Shaw. "Macroeconomics of Unbalanced Growth: Comment." *American Economic Review*, 1968, *58*(4), pp. 877-84.
Birch, J.W. and Cramer, C.A. "Macroeconomics of Unbalanced Growth: Comment." *American Economic Review*, 1968, *58*(4), pp. 893-96.
Blatt, J. "How Economists Misuse Mathematics," A. Eichner, *Why Economics Is Not yet a Science.* London: Heinemann, 1983,
Borcherding, Thomas E. "The Causes of Government Expenditure Growth: A Survey of the U.S. Evidence." *Journal of Public Economics*, 1985, *28*, pp. 359-82.
Bradford, David F. "Balance on Unbalanced Growth." *Zeitschrift für Nationalökonomie*, 1969, *29*, pp. 291-304.
Bradford, David F.; Malt, R.A and Oates, W.E. "The Rising Cost of Local Public Services: Some Evidence and Reflections." *National Tax Journal*, 1969, *22*(No.2), pp. 185-202.
Friedman, Milton. *Dollars and Deficits*. Englewood Cliffs, NJ: Prentice-Hall, 1968.
Hazlitt, Henry. *The Failure of the "New Economics"*. Princeton, NJ: Van Nostrand, 1959.
Keren, Michael. "Macroeconomics of Unbalanced Growth: Comment." *American Economic Review*, 1972, *62*(1/2), pp. 149.
Keynes, John Maynard. *How to Pay for the War*. London: Macmillan, 1940.
Lybeck, Johan A. *The Growth of Government in Developed Economies*. Aldershot, UK: Gower Publishing Company, 1986.
Lynch, L.K. and Redman, E.L. "Macroeconomics of Unbalanced Growth: Comment." *American Economic Review*, 1968, *58*(4), pp. 884-86.
Mueller, Dennis C. *Public Choice Ii*. Cambridge: Cambridge University Press, 1989.
Oulton, Nicholas. "Must the Growth Rate Decline? Baumol's Unbalanced Growth Revisited," London: Bank of England, 1999, 1-52.
Robinson, Joan. "Macroeconomics of Unbalanced Growth: A Belated Comment." *American Economic Review*, 1969, *59*(4), pp. 632.
Samuelson, Paul A. and Nordhaus, William D. *Economics*. New York: McGraw-Hill Irwin, 2001.
Spann, Robert M. "Rates of Productivity Change and the Growth of State and Local Governmental Expenditures," T. E. Borcherding, *Budgets and Bureaucrats: The Sources of Government Growth.* Durham, N.C.: Duke University Press, 1977, 100-29.
Stiglitz, Joseph E. *Economics*. New York: Norton, 1993.

Worcester, Dean A. Jr. "Macroeconomics of Unbalanced Growth: Comment." *American Economic Review*, 1968, *58*(4), pp. 886-93.

---

[1] An internet search performed on July 25, 2009 found 4,800 links for "Baumol's Disease", 1,400 links for "Baumol effect" and 325 links for "Baumol's Curse".

[2] Because Baumol's entire arguments rests upon the increased productivity of labor in the progressive sector, it might be regarded as a somewhat singular simplifying assumption that labor is the only input to be considered in the formal analysis, ignoring the capital cost of implementing the productivity-enhancing technology.

[3] Note that all short-run profiteering and price behaviour may be disregarded in this analysis which essentially pertains to a long-run scenario.

[4] In standard Samuelsonian pedagogy (e.g., Samuelson and Nordhaus, 2001, pp.152-155) the long-run aggregate supply curve of a competitive industry is defined only in a scenario wherein the technology of supply remains fixed while the market demand varies. Such pedagogy does not provide a microeconomic foundation for Baumol's (1967) scenario in which market demand remains constant while the technology of supply (productivity of labor) improves exponentially. Nonetheless, the microeconomic foundations of Baumol's scenario are easily definable in that the long-run aggregate supply curve for the industry under these circumstances cannot be anything other than the long-run market demand curve. This result is also recognised in the textbook by Stiglitz (1993, p.122, Fig.5.10A) for the case of fixed market demand.

[5] While the watch industry is offered here as an example, following the example used by Baumol *et al*. (1985), we are not imputing that Baumol's (1967) original derivation was conceived in terms of a given single product or industry. We take the original concept, and the present analysis, to be in terms of a basket of technologically progressive industries that, taken as a whole, are assumed to have a unit elastic long-run market demand curve as they move through successive generations of exponentially growing technological progress.

[6] Of course, one does not need to postulate, for example, that the demand curve has a certain fixed elasticity over its quantity domain and that this curve remains fixed in the long run. Alternatively, one might postulate that, whatever the true instantaneous shape of the demand curve might be, market demand changes in such a way as to trace out a time-series of market equilibria that defines a *de facto* long-run demand curve of the supposed price elasticity. This draws attention to the intractable difficulty, well described by Hazlitt (1959, pp.101-104), of distinguishing movements along a demand curve from changes in demand in the real world.

[7] For simplicity we shall ignore the effects of prevailing winds or air currents on the dispersal of this domestic soot.

[8] While the present analysis directly contradicts the first-order dependence between sootfall experienced per household and housing density suggested by Baumol (1967), an anonymous reviewer has suggested an alternative analytical model that would agree with Baumol's suggestion. Suppose that the amount of soot generated per household is S but that this is distributed evenly over a wide area A, equal to many times the area of a house block, a. Then a solitary house located in a neighbourhood comprising a single household would receive Sa/A sootfall , whereas households in a two-household neighbourhood would each receive $2Sa/A$ sootfall. Similarly, households in an n-household neighbourhood would each receive nSa/A sootfall; i.e., the relationship would be linear in household density provided that the toal area occupied by all households remained a small fraction of the total area *evenly* covered by the soot generated by any single household.. While such a model conforms to Baumol's (1967) suggested first-order dependence, we do not regard its conditions as being sufficiently realistic to offer a valid alternative to the present analysis represented in Figure 2. Moreover, such a model would still not lead to a *second-order* relationship between aggregate sootfall and population density claimed by Baumol (1967).

[9] Strictly speaking, a passenger cannot be the person in control of the vehicle. Yet the number of man-hours of productivity lost from an economy through traffic congestion should logically include those lost by the drivers of the vehicles, especially as a large proportion of dense peak-hour traffic comprises vehicles occupied only by their drivers.

[10] Note that the *primary* relation between duration of delay and traffic density might well be higher-order, but that is not the point at issue here.

[11] It is not the purpose of the present paper to review or assess any empirical studies of "Baumol's Disease" beyond noting that such studies implicitly accept without question the validity of the theoretical arguments developed by Baumol (1967). Therefore, these studies are not relevant to the theoretical objections raised to

those arguments in this paper. The reader who wishes to peruse the empirical literature might well begin with the articles by Bradford et al., (1969), Robert M. Spann (1977), Baumol et al., (1985) or Borcherding (1985). A useful perspective is provided in a Bank of England working paper by Nicholas Oulton (1999).