Are Unrealistic Assumptions/Simplifications Acceptable?

Some Methodological Issues in Economics

Yew-Kwang Ng
Winseumius Professor, Division of Economics, Nanyang Technological University,
14 Nanyang Drive, Singapore 637332. Phone: 65-67906785; email: ykng@ntu.edu.sg

[A paper presented as a keynote speech at the 2014 Eighth Biennial Conference of Hong Kong
Economic Association at Shandong University, 13-14 December.]

Abstract

Are unrealistic assumptions acceptable? If results are not changed in a misleading
way, they are acceptable. The same assumption may be acceptable in one context and
not in another. Assuming identical pairs of individuals in the Parity and Limit
Theorems in general equilibrium theory is acceptable as results are not changed
substantially. The alleged equivalence of the first and second-price auctions is based
on misleading assumptions; Coase’s case against taxing pollution is based on a
misleading all-or-nothing comparison. The contrasting results (neutrality versus non-
nutrality of money) of perfect vs. imperfect competition in macroeconomics with
important real-world policy relevance are also used to illustrate the point.

Keywords: Assumption; simplification; methodology; Coase; Friedman.

JEL classification: B41. Economic Methodology.

---

1 I am grateful to Nanyang Technological University for research funding and to Yi Xin for excellent research assistantship.
Simplifying, often very unrealistic, assumptions are heavily used in economic analysis. More than sixty years ago, Friedman (1953) published his methodology of positive economics. One of his main points is that assumptions used in economics need not be realistic; as long as the conclusions (especially predictions) stand up to the test of empirical verification, unrealistic assumptions are acceptable as simplification for analysis. This sparked a heated debate involving prominent authors including Nagel (1963), Samuelson (1963), Machlup (1964), Coase (1975), Hausman (1989, 1992) and others, and with continued interest through the decades until recently (e.g. Mäki 2009b, 2013, Leschke 2012, Pfleiderer 2014, Gilboa et al. 2014). The passing away of Coase in September 2013 and the memorial piece by Cheung (2013) provide stimulation for reconsideration half a century after the debate.

The old methodological debate sparked by Friedman (1953) is largely unsettled. In fact, at least in accordance to some view (Mäki 2009a), Friedman’s paper itself contains ‘ingredients … hard to reconcile with one another’ (p.90) and could be interpreted ‘as a realist … in contrast to standard textbook positivism’ (p.91). In fact, Mäki wants to rewrite Friedman (1953) into a ‘realist manifesto’ (p. 113) that equals ‘realism + fallibilism + social epistemology. According to this methodology, excellent economists are *epistemically ambitious* in seeking deep truths about underlying causal structures, and at the same time *epistemically modest* in acknowledging the possibility of error and the shaping of inquiry by academic and other institutions – and again *epistemically ambitious* in demanding the design and implementation of institutions that are adequate for enhancing the acquisition of significant truths’ (Mäki 2009a, p. 114; italics original). Leaving aside the question how similar is the rewritten F53 (Mäki’s term for Friedman 1953) to the original, I largely agree with this realist manifesto, though recognizing that it is very difficult to achieve. In a sense, this paper may be regarded as a modest attempt to contribute a little to Mäki’s rewriting. It may also be taken as sympathetic to Pfleiderer’s (2014) argument that, when theoretical models are applied to the real world, they should pass through a realism filter. This convincing argument is based partly on the practical difficulties of having conclusive tests on the predictive power of all relevant models and partly on the fact that our Bayesian prior of the real world is not empty.

---

2 On the influence of Friedman (1953) on the economics profession, see Mayer (2009); on opposing views on the influences of Friedman on the emergence of formalism, see Hands (2009).
While not attempting to resolve all relevant issues, this paper attempts to address an important aspect, whether unrealistic assumptions/simplifications are acceptable. An eclectic position is argued that unrealistic assumptions may or may not be acceptable depending on the context. Economics is involved with explanation, prediction, and normative evaluation such that the deductivism of the classical and neoclassical economists (especially Mill; see Hausman 1989), the realism of Samuelson (1963), and the falsificationism of Popper (1972) and the predictionism of Friedman (1953) may have some elements of truth/acceptability but each of them is not sufficient/complete on its own. Also, while empirical verification is important, there are serious limits of empirical inference without theory, as recently emphasized by Wolpin (2013), though there are also limits of inference with theory (Rust 2014). Theories are useful if used appropriately, but not without limits. Again, some eclecticism is appropriate.

Recently, Colander (2013) contends that the economics profession's primary goal should be achieving better real-world outcomes, and its secondary goal should be better understanding of the economy for the sake of understanding, contrary to the reverse emphasis of most professional economists these days. I largely agree with this restatement of the Pigovian position that bearing fruits is more important than shedding lights. However, due to the increasing division of labor, we need not preclude a large fraction of professional economists being concerned mainly in shedding lights, provided these are related to the real world and hence may be fruit-bearing one day. This makes realism, at least in terms of the relevance of results, important, though some obviously unrealistic assumptions may yet be acceptable as harmless simplifications, or even useful simplification to allow focus on the important relationships.

Taking an eclectic position, I also largely agree with the standing of Gilboa, et al. (2014, pp. 5-6 of paper) that ‘Viewing economic models as explorations [Hausman 1992], isolations [Mäki 1994], capacity-identifiers [Cartright 1998], and credible worlds [Sugden 2000, 2009, 2010] are … not mutually exclusive … We share many of these views’. Even the caricature of economic theorists’ dilemmas by Rubinstein (2006) has some elements of truth including the ‘fable’ aspect economic models.3

3 However, I do not find the dilemmas discussed really dilemmas. Consider the first dilemma. Assuming that an apple any day is always preferred to 2 apples the next day (assuming certainty for both), Rubinstein shows that, one must logically also prefer a single apple now to receiving one every day from day 18 onwards for the rest of one’s lifespan of 120 years. That the latter is obviously absurd constitutes the dilemma. In my view, it is the absurdity of the former
On the important question of the acceptability of unrealistic assumptions, the position of this paper is that this depends on whether the unrealistic assumptions concerned change the results in a significant way so as to be misleading. Section I provides some simple examples where the answers (one negative and one positive) are clear. The alleged equivalence of the first and second-price auctions is based on unacceptable assumptions. On the other hand, the use of identical pairs of individuals in the Parity and Limit Theorems does not change the conclusion in a misleading way. Section II argues that Coase’s case against the Pigovian tradition of taxing pollution (in the presence of transaction costs, not the Coase’ theorem is the absence of transaction costs) is based on his simplification of all-or-nothing comparison that led him and others to ignore an important asymmetry. Section III argues for the importance of preferring more realistic assumptions where manageability and fruitfulness are not too seriously compromised, using the example of perfect vs. imperfect competition at the macroeconomic level. Section IV concludes. Instead of using mathematics or complicated methodological discussion, this paper mainly uses simple examples to illustrate points made, in such a way that virtually all economists and students should be able to understand the main argument.

I. The Acceptability of Unrealistic Assumptions: Some Contrasting Examples

Whether simplifying assumptions are acceptable depends on the context. If they simplify the analysis without changing the conclusions substantially, they are acceptable; if they make the conclusions misleading, they are not acceptable. Let us consider some simple examples, both hypothetical and actual important analyses in economics.

Suppose we want to calculate the amount of time it takes to drive from one location A to another B hundreds of kilometers apart by driving a car at a constant (in fact, only approximately constant or average) speed. In the calculation, we assume that the car is a single point (ignoring its length). We just divide the distance between A and B by the speed of travel and get the conclusion that it will take \( x \) hours. Obviously, for such an exercise, the simplifying assumption that the car is a single point is an acceptable simplification. The conclusion gives us the right answer we want.

---

preference that fully explains the absurdity of the latter preference, as 2 to the power of 17 gives a number many times more than the number of days in 120 years. The preference for an apple now to two the next day appears less absurd because two apples tomorrow may often be tainted by uncertainty, irrational impatience and the like. For the special case where the first apple is essential for survival over the day, both preferences are not absurd and no dilemma is involved either.
Now consider a different exercise. Suppose we want to estimate the probability of a collision if we drive a car across an intersection. Given the knowledge of the frequency of cars travelling across the same intersection from the road perpendicular to the one we are driving, we may do the calculation accordingly. If we assume that cars are just points with no length, we get the conclusion that we may as well cross without looking as the probability of collision is of measure zero even in a busy intersection! The same assumption that cars are points is an acceptable simplification in the first calculation above but a terribly misleading one in this second calculation. The explanation is fairly obvious. Given the speed, the time of travel is not affected by the length of the car, but the probability of collision is much affected by the size of cars. Thus, we cannot just look at an assumption itself and reject or accept it; we have to consider the specific context in using simplifying assumptions in analysis. In the two hypothetical calculations discussed here, the difference is obvious and people are unlikely to commit the mistake of crossing a busy intersection blindly. In other words, commonsense is sufficient to guide us here. However, in more complicated or less obvious cases of economic analysis, mistakes similar to blind-crossing (at least in analysis) actually exist, as discussed in Example 2 below. But first we consider the case where some very unrealistic assumption is actually acceptable.

**Example 1** where a very unrealistic assumption does not significantly affect the conclusion but serves to make the analysis/proof simpler, and hence is an acceptable, even a good assumption.

**Parity and Limit Theorems**

In the proof/demonstration of the limit theorem in general equilibrium theory [that the core shrinks as the number of trading individuals increases and shrink to a single point/allocation which is the competitive equilibrium; see Edgeworth 1886, Debreu & Scarf 1963], a simple assumption is used that the increase in number of individuals takes the form of identical pairs of individuals. Identical individuals are defined as those with identical endowments of goods/resources and identical preferences [indifference maps]. This allows the parity theorem [identical individuals must end up with identical allocations for the allocation to be in the core of the economy] to be used, making the demonstration much simpler. After the demonstration, it is clear that the reason for the shrinking of the core is due to the increase in trading options. Then one can sees that, even if the increased
individuals are not identical, the increase in trading options still applies in most cases [some contrived special examples where the new individual only has some special ‘goods’ where he adores and no existing individuals want are possible; the core then does not shrink]. Then one can see clearly that the assumption of identical pairs of individuals is innocuous. It simplifies the demonstration without significantly changing the conclusions. It is a good assumption.

Similar to the use of identical pairs of individuals in Parity Theorem is the simplification of using a representative agent (usually an individual or a firm) in many analyses. For problems not focusing on distributional issues, this abstracts away not relevant differences between individuals and concentrates on the common elements. However, one has to be careful in using this simplification, including not attributing what all agents can do together to a single agent, even if representative. (See Ng 1986, Ch. 2 and Ng 2007a.) Appropriately used, this simplifies the analysis without getting misleading results, though additional results from individual differences or distributional aspects may be lost and have to be analyzed separately. In particular, Ng (1986, Appendix 3I) uses a fully general equilibrium analysis to show that (1) for any (exogenous) change (in cost or demand) there exists (in a hypothetical sense) a representative firm whose responses to the change accurately (no approximation needed) represent the response of the whole economy in aggregate output and average price, and (2) a representative firm defined by a simple method (that of a weighted average) can be used as a good approximation of the responses of the whole economy to any economy-wide change in demand and/or costs that does not result in drastic inter-firm changes.

Another similar example is Xiaokai Yang’s simplification that all individuals are identical ex-ante in his analysis of the division of labor. (See Yang 2001, Yang & Ng 1993.) This simplifies the mathematical manipulation and allows many insightful results to be obtained. [These results are so important that Buchanan (2004) twice nominated Yang to the Nobel Prize Committee before the sad passing away of Yang at the age of 55 in 2004.] Moreover, it also allows the focus on the economies of specialization (from the division of labor), showing more clearly that gains from trade may be achieved without the existence of exogenous comparative advantage. This is again an acceptable simplification provided one does not conclude that comparative advantage does not exist in the real world.
Example 2 where the result is drastically changed and the assumption is not acceptable.

**Auction Mechanisms:** What method of auction will ensure that the item goes to the one who values it highest and possibly yields a high price for the auctioneer?

First vs. second price sealed-bid auctions. The latter is the Vickrey auction which ensures that each individual bids to the true valuation (maximum willingness to pay) and hence ensures that the one with the true highest valuation wins the item.

In the common first-price sealed-bid auction, the one who bids highest win the item and pay the price bid. Under such an arrangement, if one bids the highest willingness to pay, if she ‘wins’, the net win after paying the price is actually zero. Thus, there is no point in winning unless the price bid is less than one’s maximum willingness to pay. It is clear that this arrangement will not elicit bidders to bid their maximal values. As different bidders try to hide their maximal values by different extents, the item may obviously go to the highest bidder who does not value the item highest. The first-price sealed-bid auction is, in general, not efficient.

The second price sealed-bid auction still awards the item to the highest bidder but charges her only the price bid by the second highest bid. This simple but ingenious twist of Vickrey makes much difference. Any (independent, no collaboration) bidder then has no incentive to hide her maximal value (ignoring some irrational/behavioral elements that may complicate the situation). If she underbids, she may regretfully lose to a higher bid. There is also no point in underbidding, as one does not pay the price bid but the price of the highest losing bid. Thus this Vickrey auction motivates all independent bidders to bid their maximal values, ensuring that the item goes to the one who values it most. It is efficient.

The above is simple ABC in auction theory. It is thus surprising to find a ‘proof’ on the equivalence of the first and second-price auctions published in reputable journals (surveyed by McAfee & McMillan 1987, Section V) and reproduced in arguably the top textbook in
microeconomics. In microeconomics circles, it has been said (with my approval) that, if one read and understand just 3 books (McCloskey 1985, Varian 1992, and Mas-Colell et al. 1995), one practically knows most if not all microeconomics (certainly more than what the present writer knows). As Mas-Colell et al. 1995 is at the highest level of the trio, it may be taken as the top micro text.

The essence of the proof (Mas-Colell et al. 1995, Section 23B, p. 862-6) of equivalence of the first and second-price auctions may be demonstrated with a simple example of just two bidders competing for a single house. Bidder A values the house at (maximal value) $300k and B values it at $280k. Two crucial simplifying assumptions (in addition to a more acceptable one of risk neutrality) are made:

1. That their estimates of the lowest possible price the other bidder would bid are the same (zero);
2. The probability distribution of possible prices bid by the other bidder from this common lowest level to the maximal value of this bidder is uniformly distributed.

Possible prices beyond this level are not relevant as a bidder is not interested in buying if the price bid by the other bidder is beyond her maximal value. Thus, bidder A assumes that B may bid any price between zero and $300k with equal probability. (A does not know that B’s maximum value is only $280k.) What price should A bid to maximize the expected net gain? Under the simplifications assumed, the (expected) probability of winning increases linearly as the price bid increases, but the (expected) net gain of winning also decreases linearly. Thus, if we plot in Figure 1 the net gain on the horizontal axis and probability of winning on the vertical axis, the curve relating these two variables is a downward-sloping straight line. Within the range A is willing to bid (not over $300k), the winning probability is highest if A bids at $300k and it decreases linearly towards zero as the price bid decreases to zero. The expected net gain is measured by the rectangle formed by a point on this downward-sloping straight line. For example, if A bids the price of $150k (midpoint between zero and $300k), the expected net gain is reflected by the size of the rectangle in Figure 1. Obviously, the rectangle of expected net gain is largest if the price bid is at this midpoint of $150. At any other point (price bid), the percentage increase in one variable is less than the percentage decrease in the other, making the product smaller. Thus, maximizing the expected net gain, A bids the price of $150k. Similarly reasoning shows that B bids also the midpoint (between zero and her maximal value of $280) price of $140k. The item then always goes to the bidder who values it highest. The result is the same as the Vickrey or second-price auction. Q.E.D.
The above demonstration is the essence of the alleged ‘proof’ of the equivalence of the first and second price sealed-bid auctions. However, the result is not applicable to the real world. The relaxation of the simplifying assumption that the lower bound of (equal) distribution of the possible bid of the other individual is the same for both individuals destroys the result. If B’s valuation is $280k and estimates that A could put in a bid for the house from $260 upward to more than $280k (how much more is not important, since B is not interested in buying at more than $280k), with equal probability distribution between $260 and $280, B maximizes her expected net gain by bidding $270k. Suppose A values the house at $300k but estimates that B may bid for the house from $220 to more than $300k, again with equal distribution over this range. A then bids $260k and B wins the bid. The house goes to the wrong person.

One may argue that the zero point should be accepted as the common lowest possible bid. However, even if we maintain this assumption, if we relax the other assumption [bulleted point 2 above] that the distribution of the probability of price bid is uniform throughout (which may be relaxed to be equal proportional between the two bidders, still too restrictive), the equivalence result is also no longer true. If a house is worth around the range of $250-300k, it is obviously not true that the probabilities of bids at the range of a few dollars could be the same as that around $200k+. Obviously, each bidder will form a rough estimate of the non-uniform probability distribution of the
other. It is possible that the bidder with a higher valuation may estimate a lower probability-weighted average bid of the other bidder than vice versa and thus bid lower that the other bidder. The item could still goes to the wrong person. To ensure equivalence, both the bulleted assumptions 1 and 2 above are close to being essential. But taken together, these assumptions are not only very unrealistic but the actual situations in the real world may give results (non-equivalence) significantly different than those derived under the unrealistic assumptions. Such simplifying assumptions, at least in yielding the equivalence result, are not acceptable. If we accept the ‘case-based’ view of economic models of Gilboa et al. (2014), models using such assumptions regarded here as unacceptable are not wrong, but their similarity to the real world is very low and are thus not very useful.

What simplifying assumptions are acceptable depends much on the context and the purpose of the analysis. Usually, some simple commonsense allows us to make the appropriate judgment. The next example shows that the same assumption (perfect divisibility) may be either acceptable or unacceptable depending on the context.

**Example 3: Perfect Divisibility**

In economic analysis, it is convenient to assume that the relevant economic variables (goods and services, factors of production) are perfectly divisible. This allows us to illustrate the analysis in Euclidean space (typically in simplified two-dimensional figures), use calculus, and put the relevant optimality conditions in simple and exact forms such as the equality of the marginal rates of substitution with the price ratios. It is true that physical goods like sugar and water are made of molecules. They are thus not perfectly divisible. If you divide water to the level of a single molecule of H₂O, further division would make it no longer water, but it will becomes hydrogen and oxygen atoms. Going further, things are quantum and hence also not perfectly divisible. However, the non-perfect divisibility of goods does not really affect the acceptability of such results as downward-sloping demand curves and $MV_x = P_x$ (consuming until the marginal value of a good equals its price) for most fairly divisible goods like most groceries and many other items. The indivisibilities at the molecular and quantum levels are too small to matter in everyday consumption, though the purchase/renting of a house or an apartment may be much less divisible. However, instead of
buying/renting a larger number of houses/apartments at lower prices, people may buy larger and better quality ones. The result is still essentially fine.

The same assumption of perfect divisibility, applied to all factors of production, allows the derivation of the conclusion that the degree of returns to scale in production must be constant (Friedman 2007, p.142-143). While Friedman himself recognized the existence of indivisibilities in the real world, his result makes some economists (e.g. Li 2011, p.8) claim that it is wrong to allow for increasing, constant, and decreasing returns to scale in production; only constant returns are possible. Obviously, such claims are incorrect, as indivisibilities and increasing returns to scale exist in the real world. One important kind of indivisibility is that an individual consumer and/or worker is indivisible. You cannot cut a worker into two halves and expect that they can continue to work! Due to this indivisibility, factories, shops, etc. have to be of some minimum sizes. If your shop is so small that no customer can come in to inspect your wares and buy, you have to sell to frogs or ants but they have no cash! The minimum sizes in shops, factories, machines, etc. make the fixed costs of producing even a tiny amount of output sizable and thus make the average cost curve sharply decreasing in the initial range; increasing returns exist. Thus, the same assumption of perfect divisibility may be a harmless or even useful simplification or one that leads, at least in the hands of careless users, to misleading conclusions.

II. Coase’s Misleading Simplification of Comparing All or Nothing

The reputably most cited paper in economics is Coase’s (1960) paper analyzing the problem of social cost of an external effect. It is well-known that the famous Coase’s theorem is based on the assumption of zero transaction costs for the parties concerned to negotiate an efficient solution. Especially for such problems as environmental disruption involving many individuals and the future, this assumption is clearly unrealistic. However, we cannot criticize Coase on this. Coase himself recognized that market transactions ‘are often extremely costly’ (1960, p.15). He also emphasized the point thus: “The world of zero transaction costs has often been described as a Coasian world. Nothing could be further from the truth. It is the world of modern economic theory, one which I was hoping to persuade economists to leave” (Coase 1988, pp.174-5). Let us thus consider NOT the Coase’s theorem (where transaction costs do not exist) but his main message of criticizing Pigou’s
proposal of taxing external costs, in the presence of important transaction costs. The methodological issue involved here is his comparison of the two extreme situations only, free pollution and no pollution at all. This simplification of all-or-nothing comparison misled him and his fellow Chicago economists into the mistake of being against the sensible Pigovian tax solution, at least if Cheung’s (2013) account discussed below is anything close to being true.

Nobel laureate Ronald H. Coase passed away on 2 Sep. 2013 at the age of nearly 103. His mutual admirer Steven Cheung, jointly with the School of Economics and Finance of Hong Kong University, organized an international conference in memory of Coase on 19 October 2013 attended by another Nobel laureate Robert Mundell and other luminaries. In the memorial handbook published for this occasion, Cheung (2013, p.27) has the following description of a meeting in Chicago in 1960:

“This is the most famous debate in the history of economics. The debate took place in A. Director’s home … Coase asked, ‘If the production of a factory pollute the neighborhood, should the government impose some restriction on the factory, such as using taxation or other methods to reduce the pollution of the factory?’ All [including M. Friedman, G. Stigler, A. Harberger, M. Bailey, R. Kessel, J. McGee, G. Lewis, L. Mints in addition to Director; many of whom became Nobel laureates later] on board agreed that the government should intervene – just like the current opinion on environmental protection in Hong Kong. But Coase said, ‘Wrong!’ The following debate lasted 3 hours, and the result is that Coase stood unshaken.”

Cheung (2013, p. 45-6) also recalls, “In 1970 I published ‘The structure of a contract and the theory of a non-exclusive resource’. It is a long paper but its content really only says: There is no such thing as externality, stupid! Of course, I was criticized by thousands, but today that paper is still there, and could be found in many reading lists of graduate schools. Those thousands of criticizers are now nowhere to be found! Doing research is an enjoyable game.” This is not just an extreme view of an idiosyncratic Cheung; for example, a prize-winning paper suggests that ‘we should expunge the concept of externality’ (Anderson, 2004, p. 460). Similar views are found elsewhere; e.g. Randall (1983). The persistence of such denials of the existence of external costs is difficult to comprehend; not only is global warming a serious problem, the shortening of life expectancy due to air pollution is also well established (e.g. Chen, et al. 2013).
The influence of Coase is particularly strong and widespread in China, as documented by Cheung (2013, pp. 19-20). Using the Chinese search engine Baidu, Cheung found 9.54 million citations for Coase, many times those of Friedman (2.53m), Hayek (1.59m), and Samuelson (1.18m). Another indication of Coase’s influence is this incident. An author (He, 2000) of *Economic Highlights* (an influential Chinese weekly on economic issues at the time) criticized a proposal for imposing pollution and congestion taxes on cars and petrol in China on the grounds that Coase’s analysis had been ignored. He wrote: “Obviously, [the author] is wrong. His mistake is exactly the Pigovian tradition criticized by Coase. He considers only the damage on one side, that of external costs of the usage of private cars. … But he forgets to calculate another account: the restriction of the usage of private cars results in how many losses on individual utilities, the car and other related industries in China?”

While the mistake of denying the existence of externalities (environmental disruption including greenhouse gases in particular) has been discussed elsewhere (Ng 2007b), the focus here is on the methodological simplification of making all-or-nothing comparisons only. It is probably this simplification that, at least partly, led Coase to reach the misleading, to say the least, conclusion of rejecting the appropriate Pigovian taxes on external costs and led that group of 21 prominent economists in Chicago in 1960 to accept Coase’s view, accepting Cheung’s account. Cheung’s account is supported by the Memoirs of Stigler (1985, p. 76): “In the course of two hours of argument the vote went from twenty against and one for Coase to twenty-one for Coase. What an exhilarating event! I lamented afterward that we had not had the clairvoyance to tape it”.

The problem of this methodological simplification may be shown using the textbook analysis of an external cost in Figure 2. The horizontal axis X measures the amount of some activity/pollution/emission that yields a (net) marginal value (MV) to the polluter/emitter measured by the downward-sloping curve AEP (need not be linear). In the business-as-usual or no restriction option, the polluter undertakes X until the private optimal point P where the net MV equals zero. If

---

4 That Coase was making an all-or-nothing comparison may be seen: ‘Assume that a factory which emits smoke … causing [total] damage valued at $100 per annum. Assume that the taxation solution is adopted and that the factory owner is taxed $100’ (Coase 1960, p.41; the word in square brackets is added by the citer). Note that both the damage and the tax are in single sums in the total rather than varying in accordance to the amount of pollution.
this activity X imposes no damages on anyone (and ignoring other problems like ignorance and second best), this private optimal point is also the social optimal point. However, if it imposes damages on some others or the whole society represented by the marginal damage curve BEC, the social optimal point is where the two curves intersect at E where the level of X is at S. Given the damage imposed, the point P is not efficient as the amount of pollution over the range SP yields benefits of only triangle ESP, but results in damages of a much bigger area ESPC. The Pigovian solution is to impose a tax on this external cost on the polluter to shift the marginal value curve from AEP downward to pass through the point S, motivating the polluter to reduce the level of X from P to S. Using the method of all-or-nothing comparison, Coase criticize this proposal as discussed below.

Coase criticized the Pigovian tradition for ignoring the reciprocal nature of the problem and concentrates only on the restriction of the polluters’ activity that harms the society. But this restriction harms the polluters. The problem is to avoid the more serious harm. A change of approach is necessary. As Coase did not use either maths or graphical analysis, he compared only the free-to-pollute situation P with the no pollution situation at the origin (Coase, 1960, p. 41). For such an all-or-nothing comparison, Coase correctly concluded that either situation may be more efficient than the other and a case-by-case examination is necessary. In terms of Figure 2, this depends on whether triangle ABE is larger or smaller than triangle EPC. As a general conclusion is not possible,
the presumption against external-cost causing activities or in favour of restricting pollution is not warranted. This apparently valid deduction misled both Coase and his Chicago debaters.

Though Coase’s conclusion is valid for his all-or-nothing comparison, it is not valid against Pigou. Pigou and his followers did not advocate prohibiting pollution outright or eliminating external costs all together. Rather, Pigou just proposed the imposition of taxes on external costs to reduce their levels towards the social optimal ones, or from point P to point S in terms of Figure 2. This is an efficient change as, at the free pollution point P, the net marginal value to the polluter is infinitesimal, while the marginal damage to the sufferers of pollution is large (measured by PC in Figure 2). This asymmetry in marginal value was missed by Coase in his all-or-nothing comparison which naturally focuses on the relative size of the two triangles ABE and EPC. Coase was only correct against the extreme environmentalists who want to ban pollution altogether but not correct in arguing against Pigou which was the main objective of that 1960 paper. [Pigou was explicitly, prominently mentioned on the very first page of Coase’s 1960 paper. Section VIII of the paper is entitled ‘Pigou’s treatment in “The Economics of Welfare”’; the second last section or Section IX is entitled ‘The Pigovian tradition’ and criticized and the last section is entitled ‘A Change of Approach’ (i.e. from the Pigovian tradition of taxing external costs).] Admittedly, the Pigovian case for taxing pollution should be taken only as prima facie, since administrative and other costs of taxation should also be taken into account, not to mention the danger of letting an inefficient and possibly corrupted government having more excuses to raise revenues. However, even if the environmental scientists’ prediction on the likely catastrophic outcomes of the business-as-usual option only has 5% probability of being realized, we certainly should address the problem seriously. Even if a report of a time bomb in my flight has only 1% reliability, I certainly want to change my flight!

III. The Adequacy of Prediction Verification

While the verification of prediction is very important, it may not be adequate. Consider the theory that the sun circles around the earth. This theory predicts that we will observe the sun rising every day. If we confine to the verification of this prediction, 100% confirmation will be obtained. But we know that this does not ensure the acceptability of the theory. If our earth does not have a moon and if stars are not visible due to some reasons, we may have difficulties discovering the falsehood of this theory.
Friedman (1953) did not rely purely on the verification of predictions. He noted that there are many different theories consistent with observations. Additional criteria like ‘simplicity’, ‘fruitfulness’ (wider applicability) may be relevant and some degree of arbitrariness may be involved. However, he placed much less emphasis, if any, on the importance of the realism of assumptions than many others including Samuelson (1963) and Pfleiderer (2014). An eclectic position is argued in this section that the realism of assumptions itself is a virtue, though we often have to trade-off with manageability and other desiderata.

Here, we use the example of the assumption of perfect vs. imperfect or monopolistic competition referred to by Friedman.

‘The theory of monopolistic and imperfect competition is one example of the neglect in economic theory of these propositions [on the unimportance of the realism of assumptions]. The development of this analysis was explicitly motivated, and its wide acceptance and approval largely explained, by the belief that the assumptions of “perfect competition” or “perfect monopoly” said to underlie neoclassical economic theory are a false image of reality. And this belief was itself based almost entirely on the directly perceived descriptive inaccuracy of the assumptions rather than on any recognized contradiction of predictions derived from neoclassical economic theory.’ (Friedman 1953, p. 9; words in square brackets are added).

Friedman might be correct that, by 1953, there had been no recognized contradiction of predictions derived from perfect competition. Whether a firm is a perfect or an imperfect competitor, its demand curve for an input is downward sloping, an ad valorem or per-unit tax on its product is predicted to increase its price. Such common predictions may be verified. Verification of predictions up to 1953 then did not differentiate between perfect and imperfect competition. Friedman opted for simplicity and rejected the latter as unnecessarily complicated. Here, we wish to argue that, provided that the complications involved are not unmanageable, a case for more realistic assumptions may be made, on the ground that perhaps other predictions of a more realistic theory may stand the test of time better. This is in fact the real situation regarding the assumptions of perfect vs. imperfect competition. While there may be no significant differences in the prediction of the two theories regarding the two microeconomic problems (input demand and taxation) mentioned above, there is a big difference when applied to the macro picture; there are also other different predictions as well at the micro-level which favour imperfect competition (Ch. 5 of Ng 1986). Introducing imperfect or non-perfect competition to macro analysis, Ng (1977, 1980, 1982, 1986, 1992) show that some big
differences in prediction exist and the ones based on non-perfect competition are more consistent with events like the global financial crisis and the persistence of double-digit unemployment in many countries over many years.

Under the assumption that firms are perfectly competitive, a central result in modern economics is the neutrality of money. This says that, ignoring possibly very short-term effects due to time lags, money supply affects only the price level, not real economic variables like aggregate and relative output levels, employment levels, and relative prices. The neutrality of money may be illustrated with the simple diagram of Figure 3. As all firms are assumed to be perfectly competitive, their situation may be illustrated by one representative one facing a horizontal demand curve for its product at the market equilibrium price \( p \). It can sell any quantity it likes at this price which thus is also its marginal revenue curve. It maximizes its profit by producing and selling at the point \( q \) where its upward-sloping marginal-cost curve cuts the horizontal demand curve. At this equilibrium point, the firm does not want to sell more. (One will be hard put to find firms in the real word that do not want to sell more at the same prices. Most firms like to sell more; many incur heavy advertising/promotion costs to sell more. They do not sell more only because they have to decrease prices to sell more. If so, they face downward-sloping demand curves. This shows the rather unrealistic nature of perfect competition.) Confining first to the short run (defined by the given number of firms), the output of the firm (after multiplication with the given number of firms) also indicates the aggregate output of the economy and the price indicates the price level.

Now consider an increase in money supply that increases the nominal aggregate demand. Since firms do not want to sell more at the same prices, this higher demand pushes prices higher. For the firm illustrated in Figure 3, its demand curve moves upward to \( d' \). As this firm represents all firms in the whole economy, this means that the price level \( P \) also increases by the same extent. In the absence of time lags, this increases the costs faced by the firm by the same proportion, shifting its marginal-cost curve from \( MC \) to \( MC' \). This is so since labor supply is in accordance to real instead of money wage rates (no money illusion). Thus, the new profit-maximization equilibrium occurs at the same vertical line or the same output level. The output level is unchanged; only prices increase. Money is neutral. This is the microeconomic foundation for the macroeconomic result of the neutrality of money.
Even if we retain all other simplifying assumptions like no time lags and no money illusion, the relaxation of perfect competition alone destroys the (necessary) neutrality of money. Money may still be neutral but is no longer necessarily neutral. That money may still be neutral with non-perfect competition is illustrated in Figure 4. (Ignore the two AC curves first.) Using ‘non-perfect competition’ is preferred to ‘imperfect competition’ as the former may subsume other market structures like monopolies and oligopolies. The representative firm is initially at the equilibrium point A with MC cutting MR from below. An increase in nominal aggregate demand (as may be the result of an increase in money supply) may just shift its demand curve proportionately vertically upward, as illustrated by the demand curve d’ in Figure 4. The new equilibrium at B then involves no change in output and an increase in price.
Figure 4 Money may still be neutral under non-perfect competition

However, for this case of non-perfect competition, the opposite (to the above Monetarist case) Keynesian case of no change in price but an increase in output following an increase in nominal aggregate demand is also possible, as illustrated in Figure 5. Here, the firm is originally at equilibrium at the point A. With an increase in nominal aggregate demand, the demand curve, instead of shifting proportionately vertically upward, shifts horizontally rightward and becomes more price elastic at the same price. This makes the value of marginal revenue at the same price (point B at the new demand curve $d'$) higher, allowing it to intersect the higher marginal-cost curve $MC'$ at a higher output level $q'$. The case of a proportionately rightward shift in the demand curve with no change in the price elasticity of demand at B (same price elasticity at B than at A) is also possible. However, for this case to be a sustainable new equilibrium with no increase in price, the value of the new marginal cost at $q'$ must not be higher than the original marginal cost at $q$. If the marginal-cost curve does shift upward in response to a higher aggregate output level (as the case illustrated in Figure 5), the marginal-cost curve has to be downward-sloping to offset the positive shift in the whole curve, to ensure that the price does not increase. A higher aggregate output level may push up input prices and hence affect the costs of firms. Taking account of the conditions of cost and demand conditions of firms in the whole economy and their interaction, I have shown (Ng 1982, 1986, and 2014, especially the updated mathematical appendix therein) more rigorously that there are plausible sets of conditions where this and/or other cases discussed above and below can prevail and persist as the new equilibrium.
If the offset just mentioned in the last paragraph is only partial so that there is some increase in price, the possibility of a partial increase in output, as may be taken to be possible using just a graphical analysis, is misleading. The increase in price of the representative firm signifies an increase in the general price level. This will shift the demand curve of the firm vertically upward over the relevant range, leading to further increases in price. A final equilibrium is reached only at the unchanged output level with only an increase in price. (To see the complete picture more rigorously, see the mathematical appendix in Ng 2014.) However, if the cost responses and price elasticity situation are consistent with no increase in price, such as the case illustrated in Figure 5, the Keynesian case of no increase in price but only an increase in output is possible. Our analysis is a general-equilibrium one (though somewhat simplified to abstract away changes in relative prices due to the use of the representative-firm methodology; see the last paragraph of Example 1 in Section I above) taking account of the responses in both the cost side (effects of changes in aggregate output and the price level on the costs of the firm) and demand side (effects of changes in aggregate output, aggregate income/demand and the price level on the demand of the firm), as well as the interaction of the firm with other firms as a whole.

In fact, a case more Keynesian than the Keynesian case is also possible. This is the case of cumulative expansion/contraction. In Figure 5, if MC' intersects MR' to the right of q', the new profit-maximization price is actually reduced. This reduction in p and hence the price level will shift
the demand curve of the firm downward, leading to a further decrease in the profit-maximization price and an increase in the output level. Such a cumulative expansion cannot of course go on forever. When firms have no excess capacity, its MC slopes sharply upward and when the economy is approaching full employment the costs of firms respond sharply to further increases in aggregate output, making the condition for cumulative expansion or even the Keynesian case no longer applicable. However, a decrease in nominal aggregate demand may trigger a cumulative contraction. Unless input prices fall sufficiently in response to a fall in aggregate output, the cumulative contraction may be quite deep, explaining why depressions are possible.

It may be thought that the possibility of the Keynesian result of an increase in output with no change in the price level following an increase in nominal aggregate demand is only possible in the short run. For the long run, we allow free entry and exit and treat the number of firms as a variable. Adding the zero-profit condition, our analysis still shows that the various cases including both the Monetarist and Keynesian cases are still possible. The Monetarist case is already illustrated in Figure 4 with the two AC curves brought back into the picture. The Keynesian case is illustrated in Figure 6. The higher nominal aggregate demand shifts the equilibrium from point A to B with no change in price, only an increase in output. This is possible even in the long run provided that the increase in the marginal cost of the firm from both its own output (from the slope of the MC curve) and the increase in aggregate output (an upward shift of MC to MC’) is offset by an increase in the price elasticity of demand (at the point B in comparison to at A due to a higher degree of competition as the number of firms has increased), and the upward shift in the average-cost curve from AC to AC’ (from an increase in aggregate output) is offset by the downward-sloping average-cost curve.
Figure 6 Non-neutrality of money even in the long run

Is the Keynesian case more or less likely to prevail in the long run, in comparison to the short run? Our analysis suggests two opposing considerations. On the one hand, the cost curves of the firm are more likely to respond positively to the aggregate output and hence shift upward more significantly in the long run as aggregate output increases. This suggests that the Monetarist case is more likely to prevail. On the other hand, the entry of new firms (as aggregate demand increases) increases the degree of competition and makes the price elasticity of demand for the product of each firm higher in absolute value, making it less likely to require an increase in price. This makes the Keynesian case more likely in the long run.

What makes the huge difference between the case of perfect competition where money must be neutral (ignoring time lags, etc.) illustrated in Figure 3 and the case of non-perfect competition where both the Monetarist case of neutrality (Figure 4) and the Keynesian case of effectiveness (Figure 5) are possible may be explained. The crux of the difference consists in both the demand side and the cost side. On the demand side, a horizontal demand curve (as necessitated by perfect competition) cannot shift leftward or rightward, it can only shift upward or downward. However, as we are not analyzing the case of a single firm, but a firm representing the whole economy, an upward or downward shift in the demand curve or price line for the firm signifies a corresponding change in the price level. In the absence of time lags, etc., a change in the price level changes the cost (including
the marginal cost) level of the firm by the same proportion, dictating no change in real output but only a change in price, as illustrated in Figure 3. In contrast, for the case of non-perfect competition, the demand curve for the product of the firm is downward sloping. A downward-sloping demand curve may either shift (proportionately) upward and downward (with its intersection point with the horizontal axis, if any, unchanged); it may also shift rightward and leftward (with its intersection point with the vertical axis, if any, unchanged), or some combination of the two or even some irregular shifts. If the demand curve shifts vertically upward as in Figure 4, it favors an increase in price with no change in output; if the demand curve shifts horizontally rightward, it favors an increase in output with no change in price as in Figure 5. Thus, both the Monetarist and Keynesian cases are possible under non-perfect competition.

Which shift in the demand curve will be the case is not arbitrary but is determined by what will happen to the price level that is consistent also with the new cost condition in our simplified general-equilibrium analysis. Roughly speaking, if the new demand and cost situation necessitates a higher price level, it leads to a vertical shift; if the new demand and cost situation allows a higher output level with no change in the price level, it leads to a rightward shift. One further complication is that, in the former case there are only nominal changes and the homogeneity (of degree zero) nature of demand functions prevails, making the vertical shift perfectly proportionate. In the latter case, real output and hence real income levels change and hence may cause a change in the price elasticity of demand; see Ng (2014, especially the appendix therein).⁵

Another crucial difference between perfect and non-perfect competition concerns the cost side. As the demand curve for the product of a perfectly competitive firm is horizontal, a determinate profit-maximization equilibrium requires that its marginal-cost curve must be upward sloping. This means that the value of marginal cost will increase if output increases, making an increase in output with no increase in price impossible. On the other hand, for a non-perfectly competitive firm, the demand curve is downward sloping, making the marginal-revenue curve usually even more downward sloping. A determinate profit-maximization equilibrium is then consistent with an either an upward, horizontal, or even downward-sloping marginal-cost curve. This means that an increase in output needs not necessarily increase the marginal cost, making the Keynesian case possible to

---

⁵ This possible change is abstracted away in Ng (1982).
prevail. This Keynesian case is only possible but not necessarily the case, as the effect through the change in aggregate output on the costs of firms must also be allowed for.

IV. Concluding Remarks

From the various examples discussed above, it is clear that simplifying assumptions or methods of simplification may or may not be acceptable depending on the context. Exactly the same assumption may be acceptable for a certain analysis but not in another. The main point differentiating acceptance and rejection should be whether the results (conclusions, predictions) are distorted to be misleading or not. (Secondary considerations include manageability, scope of applicability, and fruitfulness.) A problem raised is: How do we make such a distinction? Or, how should we apply Pfleiderer’s realism filter when trying to use a bookshelf model to the real world? While not pretending to have a full answer to this important question (hence leaving much scope for further discussion and research in this), a couple of points may be made.

First, as already remarked above, many cases may be clearly seen by just using common sense and perhaps some simple economic intuition. This suggests that we should not purely be guided by our mathematical models and/or statistical analysis, but we should also examine our models/analysis and results with common sense. If unreasonable results are obtained, we should reconsider the models and revise the assumptions if necessary. Unreasonable results should be carefully distinguished from surprising results. The latter are those that are not obvious but yet may be applicable and provide enlightening insights not known before. A realistic model giving only obvious results is not very useful as it does not give any insights in additional to the obvious. The distinction between surprising and unacceptable results may require common sense, good intuition, and perhaps empirical verification. When starting his persuasive paper, Pfleiderer (2014) ‘was determined not to use the phrase “common sense”’, but then finds that ‘simple common sense (based on knowledge of the world we live in)’ ‘is unavoidable’.

Secondly, the usefulness of the results may be relevant. For example, the equivalence result of the first and second-price auctions discussed above gives a conclusion that misleads people not to appreciate the superiority of the second-price auction. Its usefulness is largely negative. In contrast, consider the first theorem in welfare economics which says that a perfectly competitive equilibrium is Pareto optimal under certain conditions including the absence of relevant ignorance and external effects. Not only that the assumptions of perfect competition, no ignorance and no pollution are very unrealistic, the conclusion of Pareto optimality may also be misleading if not interpreted carefully. In
fact, at least to some extent, the theorem probably has been interpreted incorrectly to support virtually 100% free-marketism (the case against the Pigovian tax on pollution discussed in Section II above is probably related to this misinterpretation). However, if interpreted properly, the first welfare theorem is a most important result in economics. On the one hand, it allows us to focus on the essence of the price mechanism or market coordination in achieving efficiency in resource allocation, production and distribution of products. On the other hand, it serves as a benchmark for us to identify the possible sources of inefficiency (e.g. monopoly, pollution, consumer ignorance) and hence derive possible corrective measures (e.g. eliminating administrative monopoly, taxing pollution, appropriate food safety regulation) to supplement the market. Thus, the unrealistic assumptions that allow the derivation of the first welfare theorem are redeemed through the usefulness (if interpreted correctly) of the conclusion.

Thirdly, we should try to be scientifically objective and not to be too much influenced by our ideological inclinations and self-interests. For example, the exclusive case for constant returns to scale discussed in Section I above could be influenced by the fact that it is related to the perfect exhaustion of the total value of output from payment according to marginal productivities of all factors involved. This perfect exhaustion is related to the acceptability of capitalism, though one may argue for Hume’s law of separating the positive from the normative. Similarly, the case against taxing pollution could be related to the inclination of many economists towards ‘big society, small government’. Though free market works in many cases, the solution of problems of external costs especially environmental disruption needs not only public actions, but also international cooperation.

REFERENCES


Mäki, Uskali (2009b), MISSing the world. Models as isolations and credible surrogate systems, Erkenntnis, 70, 29-43.


Ng, Yew-Kwang (1977). Aggregate demand, business expectation, and economic recovery without aggravating inflation, Australian Economic Papers, 130-140


Pfleiderer, Paul (2014). Chameleons: The misuse of theoretical models in finance and economics, working paper, Graduate School of Business, Stanford University.


