Analyse & Kritik 26/2004 (© Lucius & Lucius, Stuttgart) p. 346–363

Julian Reiss

Evidence-Based Economics

Issues and Some Preliminary Answers[*]

Abstract: This paper presents an outline of a methodology of 'evidence-based economics'. The question whether an economic statement is evidence-based must be answered on three different levels. The first level concerns *measurement*: it asks whether claims made about economic quantities such as inflation, unemployment, growth or poverty are justified by the data and measurement procedures. The second level concerns *induction*: it asks whether claims made about the relations between economic quantities (such as 'number of babies born predicts growth', 'change in money causes change in monetary income', 'non-borrowed reserves can be used to control the interest rate'), are justified by the inference procedures. The third level concerns *idealisation*: it asks whether the quantities and relations selected are justified by the stated aim of the inquiry. The paper provides a discussion of these three types of investigation and of some solutions that have been offered.

1. Introduction

Suppose you are a member of a central bank committee deciding on whether or not to change the interest rate. It is central bank policy to target the country's inflation rate, and the measured rate is above the set target. What do you recommend?

The aim of this article is to raise a number of fundamental questions that need to be addressed if your recommendation should receive the label "evidence-based". To call a claim about an economic relationship evidence based is neither empty nor trivial. Many a (matter-of-factual) claim is supported by a wide range of things other than evidence: belief, hope, judgement, custom, myth or, perhaps more to the point, theory. Without wishing to be heretical it seems fair to say that in the social sciences (as well as elsewhere) extra-evidential considerations have played a significant role in establishing a claim more than once.

But even if researchers agree on the aim of making claims on the basis of the best available evidence, there is considerable disagreement as to what counts as the best evidence. Does evidence have to be measured? Is only quantitative evidence admissible? What role do non-economic data such as legal documents or historical and institutional facts play? Can aggregate data provide evidence in the same way as price and quantity data? What is understood by "casual empiricism" and does it meet the demand that theoretical claims be based? Is

[*] Thanks to Nancy Cartwright for valuable comments. Research for this paper has been supported by the Arts and Humanities Research Board.

it compatible with the demand that theoretical claims be based on the best evidence? What is the role of theory in collecting evidence?

According to the position defended here, evidence for a claim about an economic relationship such as the above-mentioned relationship between the interest rate and inflation comes in three stages. The first stage is concerned with *measurement*: do the associated measurement procedures measure the quantities of interest correctly and accurately? The overwhelming majority of economic quantities are not observable in any straightforward sense: inflation, unemployment, growth, household, preferences. Hence their existence and facts about them must be established on the basis of more or less involved measurement procedures, which often include a number of physical measurements as well as aggregation procedures. Naturally, not every procedure measures its target quantity equally well. But what criteria does a procedure have to fulfil if it is to count as adequate? What can constitute evidence for a sound measurement procedure?

The second stage is concerned with *induction*: do past observations of a relationship of interest justify projections onto unobserved situations? Suppose that we have observed a negative correlation between the interest rate and inflation in the past. Does that give us reason to believe that raising interest rates helps to keep inflation down? It is a well-known fact that correlation is not the same as causation. But how can causal claims be established on the basis of observed correlations if at all? Further, given that we have correctly identified a causal relationship, does this knowledge support conclusions for policy purposes?

The third stage is concerned with *idealisation*: are the idealisations made in a given scientific inquiry justified? The wealth and complexity of the world around us makes it inevitable that any scientific inquiry suppresses and distorts information. But when can we say that an idealisation is sound? What are criteria for distinguishing adequate from inadequate idealisations? In what sense can there be evidence for the goodness of an idealisation?

In what follows, I will discuss each problem set in more detail and point out an evidence-based solution as well as solutions that have been put forward that make use of extra-evidential considerations. Throughout, the positive solutions will be sufficient but not necessary for the label "evidence based". That is, the solution I discuss is good enough but there may be other solutions that do the required job just as well.

2. Measurement: Theory and Robustness

In this section I want to discuss problems surrounding the measurement of inflation.[1] Consider the right hand side of table 1 first. Here the prices and quantities exchanged of two goods are given for two years. From the point of view of macroeconomics, these numbers are as good as meaningless. In James Bogen and James Woodward's terminology, they do not constitute a 'phenomenon'—a stable and general feature of the world which may be of theoretical interest

[1] I discuss this case study in much more detail in Reiss forthcoming a.

(Bogen/Woodward 1988). Rather, they constitute 'data'—observable outcomes of measurement or experimental procedures by means of which phenomena are established.

Information about inflation is extracted from the data using an index number formula. On the right hand side of the table the results from five standard indices are given: the arithmetic, geometric and harmonic means and the (base-period weighted) Laspeyres index as well as the (current period weighted) Paasche index. The dispersion of the results is very large indeed. This is, first, due to the small size of the sample (only two goods) and, second, the specific behaviour of the prices and quantities. But what this numerical example displays in the extreme is true in general: the choice of the index number formula may matter significantly.

	P_{coc}	Q_{coc}	P_{clov}	Q_{clov}	I_{arithm}	I_{geom}	I_{harm}	I_{Lasp}	I_{Paa}
Year 1	100	3	50	2					
Year 2	50	1	100	3	25%	0%	-20%	-13%	40%

coc=cocoa; clov=cloves; arithm=arithmetic; geom=geometric; harm=harmonic; Lasp=Laspeyres; Paa=Paasche

Table 1: Different Price Indices

But which is the right formula? In practice, there are three competing approaches to answer that question. The first, called atomistic or stochastic approach originates with William Stanley Jevons's attempts to test the quantity theory of money. According to Jevons and his contemporary followers, an individual price change is made up of two components: a central tendency (which represents, in Jevons's work, the influence of gold) and a random disturbance (which, in Jevons's work, is due to randomly distributed effects of the 'conditions of supply and demand'). The problem to find the right index number thus consists in the problem of extracting the signal in a sea of noise.

The second approach, called functional or economic, assumes that prices and quantities are functionally related through the optimising choices of economic agents. Price changes are thus not randomly distributed but rather determined by the underlying preferences of rational consumers. The problem here is to find an index which represents a quantity that is meaningful in the sense of economic theory.

The third approach, called test or axiomatic, regards an index as an abstract mathematical object that satisfies a set of desirable properties or axioms. The problem of this third approach is to find such a set of properties which is both consistent and represents the characteristics we would normally ascribe to an index number.

Unfortunately, all three approaches are theoretically incomplete. In its simplest version, the *stochastic approach* suffers from the defect that (for all we know) price changes are not independently distributed. If this assumption is made nonetheless, the estimator of the central tendency will be biased. In more sophisticated versions, however, an estimator of the relative price change is in-

cluded. But these versions are identifiable only under 'arbitrary'—not theoretically determined—identification restrictions. The quantity of theoretical interest within the *economic approach* is the ratio between the budgets at the two different price levels *under constant utility*. This quantity is doubly underidentified. First, there are in fact two utility levels, not one: the level during the first or 'base' period and the level during the second or 'current' period. Accordingly, there are two indices. Second, one of the two budgets is unobservable because it is counterfactual. For example, if we take the utility level of the base period, the budget of the current period is defined as 'expenditure at new prices using quantities had utility been constant'. In fact, however, the utility level will have changed too. The observed expenditure will therefore not be informative about the relevant counterfactual quantity. Under certain assumptions, this index can be bounded by the Laspeyeres and Paasche indices but it cannot be measured directly. The main difficulty of the *axiomatic approach* is that many sets of desirable index properties are inconsistent—no index can therefore fulfil all properties. If one or more properties are given up, many index number formulæ fulfil the smaller set. Here, too, theoretical arguments do not support an unambiguous choice.

In addition to these theoretical deficiencies, inflation measurement is beset by practical problems. The most frequently discussed ones are the following:

- *Sampling bias.* Both households (for the quantities) as well as the goods' prices must be representatively selected. There is at least a possibility that this is conducted in a biased way. With respect to this difficulty, there are however sampling techniques available which can minimise the likelihood of this bias. The other difficulties are more principled.

- *Substitution bias.* When relative prices change, consumers react by shifting funds to the relatively cheaper goods. By their very nature indices, which can use only either base weights or current weights cannot capture this effect. It is strongest when the fixed based-weighted Laspeyres index is used (such as in the US where the weights stem from 1985).

- *Quality bias.* The quality of many goods changes continuously, in particular in modern service economies. The basic unit for all inflation measures is the price *relative* or *change*. But measuring the relative or change even for a single good presupposes that the good is available in both periods. This is not the case for many goods, especially when the periods are wide apart.[2]

- *New goods bias.* This is a version of the previous bias. Not only do goods' qualities change, new goods are also constantly introduced. There is of course a continuum between the two. It is not clear whether Windows^{XP}, say, is a new good or whether it is MS DOS with changed quality (nor

[2] There are of course attempts to minimise this bias. Hedonic indices, which measure the quality of goods along a number of dimensions, are invented to this effect, and they have been implemented in the US CPI.

is it clear in what direction the quality changed, if it did). But arguably, products such as mobile phones simply were not available 30 or so years ago. It is not clear how to incorporate new goods into an index.

- *Outlet bias.* This too is a version of the quality bias. When new distribution channels arise, consumers might shift from a more expensive form to a cheaper form such as discounters or the internet. Substitution and sampling issues also arise.

Back to our query. If we want to find an evidence-based answer to the question whether raising interest rates helps to keep inflation down, we better had an evidence-based answer to the question what the magnitude of inflation is in a given period and a given region in the first place. What would an evidence-based attitude to answering that question consist in? There are many ways to justify the use of a specific measurement or experimental technique in settling a given scientific question. In the physical sciences, often theory plays a large role (see *e.g.* Shapere 1982), further causal background knowledge (see *e.g.* Hudson 1999) and more purely empirical considerations (see *e.g.* Chang 2004). The more confidence we have in our theories, the more will we be justified in using them for constructing measurement instruments.

What situation do we face in economics in general and in inflation measurement in particular? It does not seem unfair to say that confidence in the empirical adequacy of the various theoretical strands in economics is not particularly high.[3] Let us substantiate this claim with respect to the three index number approaches exemplarily.

As mentioned above, a drawback of the stochastic approach is that some models assume that price changes are independently distributed. Other models try to mend this by adding a term of the 'relative price change' of good i vis-à-vis the central tendency. Even if we grant that the problem of relative price changes can be modelled in this way (which is denied by many in the field, especially by advocates of the economic approach), there remains the difficulty that an identifying assumption must be made. One assumption which is frequently made implies that the volatility of a good is inversely proportional to its budget share. But not even proponents of the stochastic approach believe this:

> "As β_{it} is the change in the ith relative price, specification (7) implies that the variability of a relative price falls as the commodity becomes more important in the consumer's budget. Thus the variability of a relative price of a good having a large budget share, such as food, will be lower than that of a commodity with a smaller share, such as cigarettes. This is a plausible specification, since there is less scope for a relative price to change as the commodity in question grows in importance in the budget." (Clements/Izan 1987, 341)

[3] Which does not mean that confidence in the inevitability to use economic theory is not high.

But:

> "As can be seen, the variances are not inversely proportional to the budget shares as required by [the assumption made]." (Clements/Izan 1987, 345)

The economic approach, too, faces important difficulties with respect to empirical adequacy. First, even within the rational-choice framework the question arises whose preferences we are modelling. Inflation is very obviously a social notion. So one might think that we are modelling a social utility function. However, given the problems associated with aggregating individual preferences, there are great doubts about whether such a function exists—even within the economics community. Alternatively, we could think of the preferences modelled as the preferences of a 'representative agent'. But that would imply that people have identical preferences, which again is rejected even within the community (see *e.g.* Staehle 1935).

In addition to this there are all the criticisms of the rational choice framework as a descriptive model of consumers' behaviour. It is easy enough to verify that the model is not literally true. But nor does it seem to be the case that people behave *as if* they were utility maximising—at least outside the very constrained situations of say economic laboratory experiments (see Guala forthcoming).

That the models of the axiomatic approach are false is less clear cut. This is due to the non-empirical nature of the assumptions that go into the models of this approach. If we knew what inflation was and how to measure it, we could test whether an index that does measure inflation satisfies a certain set of mathematical properties. But I do not see what in the world can be represented by the *a priori* desirable properties of an index number. But this consideration proves the point I am trying to argue for no less: namely, that the models used in the making of an inflation index do contain few (approximately) empirically adequate assumptions.

Now, there is one strategy that has often been employed in situations where confidence in any particular instrument is low (due to the fact that, say, the instrument is theoretically not very well understood): check one instrument by means of another. C.D. Broad ascribes a strategy like this to Francis Bacon:

> "The senses have two defects, one positive and the other negative. The positive defect is that there is always a subjective element in sensations; they represent things as they affect a particular organism in a particular place and not simply as they are in Nature. The negative defect is that the senses respond delicately only to a very narrow range of stimuli. They overlook what is very small or distant or swift or slow or weak or intense. Bacon holds that these negative defects can be largely overcome by the use of instruments and by other devices which he discusses very acutely in the Novum Organum under the name of Instances of the Lamp. The subjective element again can be eliminated by judicious comparisons between one sense and other and one percipient and another. The deliveries of the

> senses, when thus supplemented and neutralised, are the solid and indispensable foundation of all scientific knowledge." (Broad 1926, 52f.)

An evidence-based attitude towards inflation measurement would thus imply two things. First, that the results of one approach are checked against the results of another (likewise, estimations of the various biases should be checked against each other *etc.*). Second, belief in the results of any one measurement procedure is justified to the degree that the results agree with other procedures.

Whether we can settle debates about measurement in an evidence-based way depends on whether or not the different procedures agree to a sufficient extent. What is sufficient, in turn, depends on what question exactly we are asking. It would be sufficient, for instance, to observe that all indices yield higher numbers for the 1990s in the US than for the 1970s in order to have good reason to believe that average inflation was indeed higher—it is not necessary that the numbers agree. But if the question is what UK inflation April 2003-April 2004 was, all indices had better agree to a great extent.

Robustness checks such as this no doubt exist in economics. But to an empiricist's mind they are fare too rare, and if they occur, they often occur with lack of system. Too often do extra-evidential considerations play a role. As we have seen above, theoretical thoughts rank very highly in measurement. Aesthetic preferences regarding, say, mathematical simplicity and elegance are given consideration. Frequently also pragmatic matters beat concerns for evidence, for example, if data availability is an issue.

From the point of view of evidence, such considerations should not matter. The world is as it is, not as it ought to be or as we can know it.

3. Induction: Prediction, Explanation, Control

Suppose now that the quantities of interest are measured accurately. How do we establish relationships between the quantities on the basis of evidence? As we will see in detail in the next section, the choice of the kind of relationship to be established is largely a pragmatic matter. We might, for instance, be interested in predictive success. We then face the problem of finding quantities that act as reliable indicators for the future values of other quantities. A relationship of that kind may be called 'stable correlation'. Alternatively, we might want to find causal relationships between quantities. The corresponding aim is then causal explanation, and our problem is causal inference. Third, we might be interested in a special kind of causal relationships, namely, the kind that is stable under intervention. This would correspond to the aim of public policy making, and the inference is a kind of causal inference. (It is not pure coincidence that I mention here the classical aims of economics, *viz.* description/prediction, explanation and control; see *e.g.* Menger 1976/1871.)

The three types of relationships are not the same. Not every stable correlation is also a causal connection, and not every causal relation can be exploited for policy purposes. Conversely, although relationships which are stable under intervention are also causal relationships, causal relationships are not always

the best ones to use for purposes of prediction. Correspondingly, the methods of inference differ in the three cases. Let us here treat causal relationships exemplarily. (No suggestion is made that this type of relationship is any more important than others; I just happen to know more about it.)

In the past thirty years or so, causality has been a topic of great significance in the philosophy of science. Consequently, a lot is known about the metaphysics of causation and causal inference, much of which is relevant for the economics project pursued here. Of the great contemporary metaphysical schools of causation—the counterfactual theory, the probabilistic theory, the process theory and the manipulability theory—numbers one, two and four are of particular interest for economics.[4] Each of these theories comes along with a set of methods for causal inference. The counterfactual theory, according to which (essentially) A causes B if and only if B would not have been if A had not been, is reflected in, first, Max Weber's account of singular causal inference in history and, second, the potential outcomes approach in statistics. The probabilistic theory, according to which (essentially) A causes B if and only if A raises the probability of B in a homogenous reference class, is reflected in the Bayes' Nets methods as well as many applications in econometrics. Last but not least, the manipulability theory, according to which (essentially) A causes B if and only if we can manipulate A in order to bring about B, is also reflected in a variety of applications in econometrics, including Kevin Hoover's method of inference (Hoover 2001). Again, for the sake of brevity, let us focus here on only two of them, the probabilistic and the manipulability theories.

The probabilistic theory of causation is motivated, among other things, by the observation that many regularities fall short of the exceptionless universal regularity ideal but nonetheless appear to be causal. We believe that smoking causes lung cancer but allow there to be cases of smokers that do not contract lung cancer as well as of non-smoking cancer victims. However, it is clear that not all probabilistic dependencies have a causal explanation, nor do all cases of causation involve probabilistic dependence. The problem of 'spurious correlation', then, is to distinguish genuine causal dependence from spurious or *merely* probabilistic dependence.

The early theory by Patrick Suppes (1970) tries to solve the problem in two steps. First, the concept of *prima facie* cause is introduced: A is a *prima facie* cause of B if and only if A precedes B in time and $\mathrm{P}(B|A) > \mathrm{P}(B)$. Not all *prima facie* causes are genuine causes though: lung cancer is probabilistically dependent on yellow stains on people's fingers (and follows it in time), but cancer is not caused by the stains. Suppes then uses the concept of "screening off" (which is originally due to Hans Reichenbach). A third variable or set of variables C, which occurs before A, screens A off from B if and only if $\mathrm{P}(B|A,C) = \mathrm{P}(B|C)$. A is now a genuine cause of B if and only if it is a *prima facie* cause and there is no variable that screens off A from B.

[4] I omit the process theory here because although all economic processes require physical processes in order to exist, due to its physicalist undertones, the theory does not help in distinguishing genuine causal relationships from spurious ones. I have written on this in Reiss forthcoming b.

It has emerged later that not any variable set C will solve the problem (*e.g.* Cartwright 1983). Rather, C needs to include all other *causes* of B. In this formulation, A causes B if and only if it occurs earlier and A raises B's probability conditional on all other causes of B. The modern theory of Bayes' Nets is essentially a generalisation of this basic idea, and the econometrics technique of multiple regression is an application.

However, a host of widely discussed counterexamples troubles the theory and its methods. Not every genuine cause is also a *prima facie* cause. In cases where a cause brings about its effect via two different routes, the influences along the two channels might just cancel (and it is possible that there exists no variable at the generic level conditioning on which would solve the problem). Kevin Hoover (2001) cites cases in economics where we intervene such as to perfectly stabilise a variable. This variable is then not correlated with any other variable in the system but obviously causally related to one or more other variables. The screening off condition can fail in systems that are genuinely probabilistic and in "mixed" causal structures (the latter problem is pertinent in particular in cross-sectional studies where, say, data from different regions in which different causal structures may prevail are pooled).

The manipulability theory is motivated in part by the idea that we can test causal claims experimentally, and that we can exploit causal relations in order to bring about an effect. Suppose we are facing a nonsense correlation such as Elliott Sober's correlation between British bread prices and Venetian sea levels (Sober 2001). On the basis of the correlation alone we might be misled into thinking that there is a causal relation between them—the correlation persists (let us suppose) when all other causes are taken into account. But now imagine that we intervene to fix one variable—we freeze the bread prices in Britain, say. Now we would fail to observe a concomitant change in the sea levels variable and thus conclude that the relation is not causal.

Unfortunately, it is not quite as easy. This is, first, because we can imagine the existence of an agent who, entirely by chance, intervenes into the putative effect variable each time we intervene in the putative cause variable. When we freeze bread prices, the agent freezes sea levels. We, mistakenly, conclude that bread prices cause sea levels. Therefore, we want to make sure that our intervention is not correlated with any other cause (such as the agent's action) of the putative effect variable. But we also want to make sure that our intervention causes the putative effect variable, if at all, only through the putative cause variable. Suppose that freezing bread prices is part of a large-scale environmental legislation parcel, which as a side effect stops sea levels from rising. Again, our intervention would not prove anything. Third, we want to make sure that our intervention is not itself caused by the putative effect variable. This is a problem in econometrics because many interventions to influence economic variables are timed in response to the development of economic conditions. Any effect the putative cause variable might have can thus be blurred. Fourth, we don't want our intervention to upset the whole causal structure—as a consequence of which no observed change would be informative. This, too, is a problem in economics because legislation, which otherwise would be a nice intervention, might act such

as to influence the causal relations between the variables of interest rather than just one variable by itself.

The relevance to our topic—the evidential basis—should be clear: to provide evidence for a causal claim means to provide evidence for a particular kind of correlation as well as for the truth of a number of causal background assumptions. As before, there do exist strands in economics where evidence of precisely this kind is sought. For example, the natural experiments movement in econometrics aims at exploiting situations where an intervention of the kind just described occurs 'naturally' without explicit help from the researcher.

But, also as before, to an empiricist's mind these strands aren't strong enough. Too often causal claims are 'established' not by real experiments—controlled or natural—but rather by thought experiments—a model or hypothetical world in which the causal connection of interest is proved. This is not to claim that these models are of no use: they show that a causal connection of a certain kind is possible because it is actual in some possible world—the world of the model. But it is a long way from showing that a certain connection is possible to having reason to believe that it is responsible for the phenomenon.

In case of arguing from the truth of a causal claim in a model or hypothetical world to a real situation the problem is one of *external validity*—the problem of whether the claims proved in the study at hand can be exported to other situations. The advantage of these kinds of hypothetical studies is that the problem of internal validity—the problem of ascertaining that the claim made is true of the envisaged situation—is virtually absent. The claims are proved to be true by means of mathematical deduction. In other cases *internal validity* appears to be at stake. One such case, in my view, is Milton Friedman and Anna Schwartz's empirical study about the role of money in the economy.

Friedman and Schwartz have often been taken to try to establish that (change in) money is the principal cause of (change in) various variables that measure economic activity, such as monetary income. There is a lot of ambiguity already in this statement. What is meant by "principal cause"? This sometimes is spelled out as the sole cause, or the sole long-run cause or the strongest candidate among a set of contributing causes. To keep things simple, and to be able to make a strong or at least plausible case for Friedman and Schwartz, let us interpret the claims as meaning that money is a contributing cause to the variables of economic activity. That is, money influences economic activity but there may be other variables that do so too. What is the evidence Friedman and Schwartz provide?

The first piece of evidence is a long time-series with data about money stock and business cycles from 1867 through 1960. Friedman and Schwartz here show that money and business activity are highly correlated (money expanding whenever business expands, money contracting whenever business contracts) and that the monetary variable leads the business variable. Have they thus demonstrated that money causes business activity?

At best they have shown that money is what Suppes called a *prima facie* cause of business activity. And they know well that not every *prima facie* cause is also a genuine cause.

> "It might be, so far as we know, that one could marshal a similar body of evidence demonstrating that the production of dressmakers' pins has displayed over the past nine decades a regular cyclical pattern; that the pin pattern reaches a peak well before the reference peak and a trough well before the reference trough; that its amplitude is highly correlated with the amplitude of the movements in general business. ... We do not, of course, know that these statements are valid for pins and, indeed, rather doubt that they are but, even if they were demonstrated beyond a shadow of doubt, they would persuade neither us nor our readers to adopt a pin theory of business cycles." (Friedman/Schwartz 1963a, 48f.)

What Friedman and Schwartz call the "most convincing" piece of evidence in favour of the monetary hypothesis is that major movements in US history (*e.g.* the major inflations and contractions) displayed the following pattern:

> "The changes in the stock of money cannot consistently be explained by the contemporary changes in money income and prices. The changes in the stock of money can generally be attributed to specific historical circumstances that are not in turn attributable to contemporary changes in money income and prices. Hence, if the consistent relation between money and income is not pure coincidence, it must reflect an influence running from money to business." (Friedman/Schwartz 1963a, 50)

To me this sounds like an attempt to use 'specific historical circumstances' as interventions of the kind discussed above. Unfortunately, the evidence they present appears to fall short of being evidence for what I have called an 'ideal intervention'. Let us recap what its properties were:

1. I (the ideal intervention) causes C (the putative cause variable).
2. I causes E (the putative effect variable) if at all only through C and not via any other route.
3. Neither C nor E nor any of their causes cause I (except those causes that cause C and E on the route I $\rightarrow$ C $\rightarrow$ E).
4. I does not disturb the causal relation between C and E.

World War I is such a specific historical circumstance. In their monumental *Monetary History of the United States 1867–1960*, Friedman and Schwartz argue that the period before the US entered the war 1914–1917 constituted a 'classical gold inflation': Allied gold poured into the US for purchases of weaponry, thus the money stock inflated and was followed by an inflation in prices and money income.

Friedman and Schwartz provide excellent evidence for clause 1: the war, or more precisely, the weaponry purchases of the Allies, caused the large increase

in the money stock. It seems also correct to say that the "specific historical circumstances ... are not in turn attributable to contemporary changes in money income and prices", *i.e.* that clause 3 is fulfilled. It is at least plausible to claim that clause 4 is fulfilled, too. The general monetary constitution, introduced in 1913 by the Federal Reserve Act, stayed in place throughout the period and did not experience any major changes.[5] But what seems utterly implausible to me is the satisfaction of clause 2. How are we to believe that the Allied purchases affect prices and incomes if at all only through money? Would we not think that had the purchases not been paid with gold but rather made on the basis of credit the effects on money income and prices would be, if not the same, very nearly the same? Consider the following statement:

> "The reason for the altered emphasis is the wartime change in the character of international trade and financial arrangements. In the belligerent nations, private individuals reacting to price incentives were largely replaced as the major traders on international markets by governments controlling their own financial machinery. They exercised an insistent and pressing demand for American goods; created an excess of exports for the United States over imports; and paid for the excess during the period of U.S. neutrality by shipping more than $1 billion in gold, selling for dollars $1.4 billion of American securities owned by their citizens and transferred under compulsion to government control, reducing by $0.5 billion short-term loans by their citizens to the United States, and by borrowing about $2.4 billion in U.S. financial markets, a total of no less than $5.3 billion." (Friedman/Schwartz 1963b, 198f.).

It seems more than likely that the purchases should affect money income and prices through a channel different from the channel via money. But if that is so, this intervention is useless in its capacity to test the monetarist claim.

In this situation, then, what would be the evidence-based attitude? If the goal is causal explanation, Friedman and Schwartz need either to investigate further and come up with very good evidence that in the period envisaged the weaponry purchases could not have caused changes in money income and the other variables except through money or to suspend judgement as to which of a range of alternative hypotheses is correct. In the next section I will offer a third alternative: the method they use is appropriate if one understands their activities to be directed towards a goal different from causal explanation.

---

[5] It is, however, at least possible that the relation between money and the other variables was changed as a consequence of the war. Anticipating entry into the war could make people act differently towards the money they hold and thus disturb the relation between money and, say, income. For the sake of simplicity, though, let us grant this point to Friedman and Schwartz.

4. Idealisation: Methods and the Aims of Science

If one accepts that the world—natural and social—is complex, especially if unadulterated, and also that intelligibility is at least one important aim of the sciences, it is easy to understand that forms of abstraction, idealisation and simplification are ubiquitous in the sciences. Economics is no different in that respect: real, gendered, people are portrayed as sexless rational economic 'men', a vast array of individual transactions is represented by aggregated measures of inflation and national income, different business cycles are simplified into the 'reference cycle' *etc.* In this context I do not want to differentiate the various activities of, say, isolation, abstraction, simplification and so on, and use 'idealisation' as an umbrella term to cover them all. Any representation of a state of affairs is idealised in this sense if it differs from the way we immediately perceive it.

It is thus easy to see that idealisation is a necessary component in the manner the sciences represent the world around us. It also easy to see that not every idealisation is equally well justified. It cannot quite be coincidence that economic theory has consistently represented people as rational agents rather than as erratic fools, that physics represents planets by point masses rather than lumps of green cheese or that chemistry informs us about how pure rather than impure elements react with each other.

The aim of this section is to give an account of what it means for an idealisation to be 'evidence-based'. When do we have reason to idealise in one way rather than another? The simple but compelling idea behind this account is that an idealisation is justified if it can be shown that it serves its purpose. The purpose is, in turn, given by the aims of science sought by the investigator. At the most general level, I want to adopt a pluralist stance with respect to the aims of the sciences here. It is not the purpose of methodology to impose the methodologist's values on the scientist. He can only interpret and, in some cases, point out inconsistencies between ends and means. The ends, however, can only be set by the science itself.

Above, predictive success, explanation and control have been identified as 'classical' aims of economics as a science. As pointed out above, the three aims are largely independent, and they may pull in different directions. According to certain theories it is impossible to predict financial time series. The reasoning is essentially that if it were possible, someone would do so, act in order to exploit his knowledge and, as a consequence, falsify the prediction. But this does not imply that it is impossible, after the fact, to explain movements on the markets. Nor does it imply that it is impossible to control the markets through regulation. Likewise, it is not always efficient to try to predict economic time series using causal knowledge (see *e.g.* Hendry/Clements 1999). Rather, simpler models tend to beat structural or causal models in forecasting competitions because they are less sensitive to structural changes. Further, the Lucas critique teaches us that not all causal or structural parameters can be used for policy purposes.

Now this is true of the relations between quantities as well as of the quantities themselves. Certain indicators may be extremely useful for predicting quantities

of interest, but they themselves stand in no significant causal relations. The other way around, some variables may represent quantities that enter genuine causal relations but they might not be very useful for predictions. To give a non-economic example, arm circumference is sometimes taken as a reliable mortality predictor for malnourished children. But this does not mean that arm circumference is in any way useful to explain mortality, and surely it is not a good quantity to target in order to control mortality. And those quantities that help in explaining mortality (such as nutritional deficiencies of certain kinds) may not always be the best predictors because they cannot be measured very accurately or very readily.

If the topic of the previous two sections was the validity of widely employed methods of measurement and inductive inference, the topic of this section is the consistency of the methods with the aims of the inquiry. Any method, if applied correctly, may give perfectly valid results relative to what it was designed for but it may nonetheless be useless relative to the aims the investigator pursues. It does not make sense to either praise or blame any particular scientific activity without reference to the goal pursued in the case. In what follows, I want to offer an account of Milton Friedman's work on monetary economics such that his methods and the idealisations employed do not, as in the previous section, come out as faulty but rather as fully consistent with his overall goals.

It has been said above that the evidence Friedman and Schwartz offer in support of their claim that money is a cause of economic activity is wanting. But who is to say that this is what they were aiming at in the first place? I have discussed their work above *as if* causal inference was their stated aim because causal inference has often been taken to be their aim. But reading the work on monetary economics in conjunction with Friedman's 1953 methodology essay reveals another possibility, a possible interpretation which, in my view, has the advantage of providing the most consistent overall account of Friedman's work.

Very clearly, in this essay Friedman claims that the ultimate goal of what he calls positive science is predictive success: "Viewed as a body of substantive hypotheses, theory is to be judged by its predictive power for the class of phenomena which it is intended to 'explain'." (Friedman 1953, 8) In most cases when he uses the phrase that "hypotheses explain phenomena", he puts the "explain" in scare quotes in order to signify that hypotheses explain phenomena in a sense no stronger than the sense in which axioms explain theorems that can be derived on their basis. At least according to the methodology essay, thus, causal explanation does not seem to be his goal. This view is further confirmed by his practice of regarding the 'assumptions' of a certain theory as mere calculation devices to make predictions. Friedman explicitly says that "[i]nstead of saying that leaves seek to maximize the sunlight they receive, we could state the equivalent hypothesis, without any apparent assumptions, in the form of a list of rules for predicting the density of leaves ..." (24).

In addition, Friedman uses the well-known observation that hypotheses are always underdetermined by the 'facts' or data against which they are tested. Because observed data are always finite but the number of hypotheses invoked to 'explain' a given body of consistent data is infinite, there will always be a

choice among hypotheses. Facing a number of potential alternatives, his preferred choice is the 'simplest' and 'most fruitful' hypothesis, *i.e.* the hypothesis that requires the least amount of 'initial knowledge' is required in order to make a prediction, and which 'explains' the widest class of phenomena (10). Predictive success is thus the goal, and among predictively successful hypothesis, the simplest and most fruitful (or 'significant') is to be picked. Now, if only predictive success is the goal (and not causal explanation), Friedman does not require fully fledged causal inference in his monetary work. To the contrary, it would be counterproductive.

The phenomenon to be 'explained' in this case is the unanimous correlation between money and economic activity in US history from 1867 – 1960. That this is indeed a stable correlation is established by Friedman and Schwartz, in the first part of their "Money and Business Cycles" (32–48). This correlation can be 'explained' by a variety of hypotheses, including 'money causes business', 'business causes money', 'factor X causes money and business' (where X can be, *e.g.*, sunspots or debt or some other variable) and 'the correlation between money and business is brute'. Now, most of what happens in the remainder of the paper is an attempt to show that the first, monetarist hypothesis, is both consistent with the evidence in the different periods and is also more fruitful than the other hypotheses because though other hypotheses may work in individual episodes, none of them 'explains' all episodes.

What have they thus established? To assume that money causes economic activity not only predicts future economic activity reliably, it is also the more fruitful hypothesis: it is consistent with more episodes than the alternative hypothesis Friedman and Schwartz consider (that economic activity causes money). To make that claim, it is inessential whether the correlation comes about as a result of direct causal influence running from money to economic activity, as a result of reverse causation from economic activity in an earlier period to money in a later period or as a result of a common cause or some other causal connection. That their tests then cannot distinguish these different cases, as has been claimed above, is an invalid criticism of the work: the idealisations they make (*e.g.* neglecting the difference between a direct causal connection and a common cause structure) is fully consistent with their goal to find a reliable predictor.

Friedman and Schwartz point out that the data from some episodes are consistent with other hypotheses. But since only the monetarist hypothesis is consistent with all episodes, this is the preferred one:

> "These propositions offer a single, straightforward interpretation of all the historical episodes involving appreciable changes in the rate of monetary growth that we know about in any detail. We know of no other single suggested interpretation that is at all satisfactory and have been able to construct none for ourselves. The character of the U.S. banking system—in particular, for most of its history, the vulnerability of the system to runs on banks—can come close to explaining why sizable declines in money income, however produced, should generally be accompanied by sizable declines in the stock of

> money; but this explanation does not hold even for all declines, and it is largely irrelevant for rises. Autonomous increases in government spending propensities plus the irresistible political attraction of the printing press could come close to providing a single explanation for wartime inflation, according for the coincidence of rising incomes and rising stock of money without any necessary influence running from money to income; but this explanation cannot account for peacetime inflation, in which the growth of the money stock has reflected a rise in specie rather than in government-issued money; and it is not even a satisfactory for the wartime episodes, since prices rises in different wartime episodes seem more closely related to the concurrent changes in the money than to the changes in government expenditure." (Friedman/Schwartz 1963b, 53f.)

To summarise, there are two reasons for me to hold this interpretation of Friedman and Schwartz's work on monetary economics. The first is that they do not attempt to provide the crucial evidence which would enable them to distinguish the monetarist hypothesis from its alternatives. Were they to engage in genuine causal inference, this would appear as a straight flaw in their reasoning. But instead they aim at finding a hypothesis which is predictively successful. For this aim, the intricacies of causal inference can safely be neglected. Second, they appear to prefer a simple and fruitful hypothesis, a hypothesis consistent with as broad a range of phenomena as possible, to a detailed causal account of each individual episode. From the point of view of causal inference, there is no reason to prefer the simpler hypothesis to the more complex: there is no reason to suppose that only one or a small set of causal factors are responsible for the observed phenomena. To suppose this would be to make the substantial and, in my view, implausible claim that simplicity is a guide to truth. But if the aim is to find 'simple and fruitful' hypotheses that enable us to organise the phenomena efficiently, Friedman and Schwartz practise just what they preach (or rather what Friedman preaches in his methodology essay).

The philosophical problem of idealisation is, roughly, how to make sense of the apparent fact that a false descriptions of some phenomenon nonetheless may yield important insights. The account defended here is a broadly Aristotelian one with strongly pragmatist features. It is Aristotelian in that it distinguishes the 'essential' from the 'accidental' qualities of a phenomenon. The accidental qualities may be suppressed or distorted in order to get to the heart of the matter. But what the essence of the phenomenon is, is not determined by anything transcendental but rather by our scientific interests. In this sense, an idealisation is a good one if it distorts in such a way as to make the most efficient predictions, to reveal the causal connections or the relations which are stable under interventions.

.

5. Conclusion

The aim of this paper has been to raise some questions about evidence in economics and give some preliminary answers. Among the issues that may be raised in this context, three have been singled out as particularly important: (a) What does evidence for economic quantities and their magnitude consist in? (b) What does evidence for relations between economic quantities consist in? (c) What does evidence for idealisations in economics consist in?

The measurement of economic quantities is beset with difficulties. However, relative to the questions we ask, the difficulties may not matter. If the results of a measurement procedure are stable enough under plausible changes of the specification as to allow us to address a given question, our answer can be said to be based on the best available evidence—and that despite the fact that many of the assumptions that go into a measurement procedure are descriptively false.

Among the various types of relations between economic quantities that can be of interest to an economist, there are *causal* relations which I discussed as an example. Theories of causality double as tests for the presence or absence of a causal relation—under certain assumptions that come along with each type of test. Evidence for a causal relation thus consists in both passing the test as well as fulfilling the assumptions of the test.

Finally, I claimed that an idealisation was a good one if it could be shown that it serves the purposes of the scientific investigation at hand. I discussed three 'classical' aims of economics, *viz.* prediction, causal explanation and control—allowing there to be a variety of other aims. In the case study about Friedman and Schwartz's investigations in monetary economics we have seen what it means for an idealisation to be consistent with the aims of the investigation.

This gets us back to the to the question raised at the beginning of this paper: what do you do if asked whether the central bank should raise the interest rate? Of course, in the context of this paper it is impossible to give a substantial answer. The point of this paper was to urge that before any valuable recommendation can be given at all, one should have a sizeable amount of evidence at one's disposal: does the measured inflation rate represent what we think it does? How about the interest rate? Do interest rates stand in the relation with inflation we think it does? And is this the right kind of relation that is required according to the aim of our inquiry? Before these questions are not answered in a satisfactory, and this means *good enough* way, I'd rather recommend abstaining.

Bibliography

Bogen, J./J. Woodward (1988), Saving the Phenomena, in: *Philosophical Review 97*, 302–352

Broad, C. D. (1926), *The Philosophy of Francis Bacon*, Cambridge

Cartwright, N. (1983), *How the Laws of Physics Lie*, Oxford

Chang, H. (2004), *Inventing Temperature*, Oxford

Clements, K./H. Y. Izan (1987), The Measurement of Inflation: A Stochastic Approach, in: *Journal of Business and Economic Statistics 5.3*, 339–50

Friedman, M. (1953), The Methodology of Positive Economics, in: *Essays in Positive Economics*, Chicago

— /A. Schwartz (1963a), Money and Business Cycles, in: *Review of Economics and Statistics 45.1* (Part 2, Supplement), 32–64

— / — (1963b), *A Monetary History of the United States, 1867–1960*, Princeton

Guala, F. (forthcoming), *The Methodology of Experimental Economics*, book manuscript, University of Exeter

Hendry, D./M. Clements (1999), *Some Methodological Implications of Forecast Failure*, manuscript, University of Warwick

Hoover, K. (2001), *Causality in Macroeconomics*, Cambridge

Hudson, R. (1999), Mesosomes: A Study in the Nature of Experimental Reasoning, in: *Philosophy of Science 66*, 289–309

Menger, C. (1976/1871), *Problems of Economics and Sociology*, transl. by J. Dingwall and B. F. Hoselitz, Urbana

Reiss, J. (forthcoming a), Magic Models of Measurement or How Economists 'See', manuscript, submitted

— (forthcoming b), Comments on Paul Humphreys' "Theories of Causation and Explanation: Necessarily True or Domain-Specific?", in: C. Hoefer/J. Diez (eds.), Causalidad y Explicacion: En Homenaje a Wesley Salmon, Barcelona

Shapere, D. (1982), The Concept of Observation in Science and Philosophy, in: *Philosophy of Science 49.4*, 485–525

Sober, E. (2001), Venetian Sea Levels, British Bread Prices and the Principle of the Common Cause, in: *British Journal for the Philosophy of Science 52*, 1–16

Staehle, H. (1935), A Development of the Economic Theory of Price Index Numbers, in: *Review of Economic Studies 2.3*, 163–188

Suppes, P. (1970), *A Probabilistic Theory of Causality*, Amsterdam