Two Approaches to Reasoning from Evidence Or What Econometrics Can Learn from Biomedical Research

Abstract

This paper looks at an appeal to the authority of biomedical research that has recently been used by empirical economists to motivate and justify their methods. I argue that those who make this appeal mistake the nature of biomedical research. Randomised trials, which are said to have revolutionised biomedical research, are a central methodology but according to only *one* paradigm. There is another paradigm at work in biomedical research, the inferentialist paradigm, in which randomised trials play no special role. I outline the inferentialist alternative in broad strokes, apply it to a recent controversy in econometrics and draw some general conclusions concerning econometric methodology.

Keywords: evidence; randomised trials; econometrics; inference.

Julian Reiss is Professor of Philosophy at Durham University. He has a degree in economics and finance from the University of St Gallen and a PhD in philosophy from the London School of Economics. His main research interests are methodologies of the sciences (especially causality and causal inference, models, simulations and thought experiments, and counterfactuals), philosophy of economics, and science and values. He is the author of *Error in Economics: Towards a More Evidence-Based Methodology* (2008), *Philosophy of Economics: A Contemporary Introduction* (2013), and nearly 50 papers in leading philosophy and social science journals and edited collections.

Julian Reiss
Department of Philosophy
Durham University
50 Old Elvet
Durham
DH1 3HN
+44 191 334 6543
julian.reiss@durham.ac.uk

1. Introduction

Evidence-based policy currently resounds throughout the land (cf. Reiss 2013: Ch. 11). This is perhaps hardly surprising — as it should not only be the case that policy goals are democratically legitimate but also that policy interventions are justified as effective means to promote these goals. And how else could one justify a policy better than demonstrating that it is based on the best available evidence?

The demand that policies should be based on the 'best available evidence' is uncontroversial. It is not uncontroversial, however, to restrict the sources of admissible 'credible' evidence to the results of randomised trials and closely related quasi-experimental studies, as proponents of evidence-based policy do. Clearly, randomisation is no panacea. For many situations experimental designs are simply inapplicable. Macroeconomic questions cannot be addressed experimentally, for instance (Cohen and Easterly 2009). Randomised trials are often more costly, financially and ethically, than available alternatives (Scriven 2008). They do not, despite their proponents' claims to the contrary, guarantee the credibility of results (Cartwright and Munro 2010, Deaton 2010); worse, they can introduce new kinds of bias (Heckman 1992). To insist on evidence from randomised trials means to throw out much relevant and hard won information. Nevertheless, virtually all initiatives that aim to help policy makers make well-informed decisions about the effects of interventions in the social, behavioural and educational arenas follow the principles of evidence based policy (Miguel et al. 2014):

> There is growing appreciation for the advantages of experimentation in the social sciences. Policy-relevant claims that in the past were backed by theoretical arguments and inconclusive correlations are now being investigated using more credible methods. Changes have been particularly pronounced in development economics, where hundreds of randomized trials have been carried out over the last decade. When experimentation is difficult or impossible, researchers are using quasi-experimental designs. Governments and advocacy groups display a growing appetite for evidence-based policy-making. In 2005, Mexico established an independent government agency to rigorously evaluate social programs, and in 2012, the U.S. Office of Management and Budget advised federal agencies to present evidence from randomized program evaluations in budget requests.

Virtually all of the philosophical literature on evidence-based policy focuses on the virtues and vices of randomised experimentation.[1] This paper instead will take a look at the 'second best', the method recommended for situations 'when experimentation is difficult or impossible': quasi-experimental designs. Quasi-experimental designs are structurally identical to experimental designs but proceed without the benefit of interventions performed by the investigator. That is, they use observational instead of experimental data.

Both experimental and quasi-experimental analyses have sometimes been called 'design-based' (Angrist and Pischke 2010). Unlike traditional econometrics the credibility of whose results stands and falls with the credibility of the background theory from which empirical models are derived, analyses of this new kind are supposed to derive their credibility from their 'design' (hence the name): because the analysed data have been produced by certain — experimental or quasi-experimental — structures, study results are deemed more reliable than those of traditional econometrics.

Why should we believe that studies that draw on experimental or quasi-experimental designs are more reliable than traditional econometric studies? Defenders of design-based econometrics sometimes appeal to the success of experimental methods in biomedical research: just as biomedical research has profited enormously from downplaying traditional sources of evidence such as clinical expertise and cohort studies and stressing the importance of randomised experiments, empirical economics can benefit from doing the same.

The aim of this paper is to look at this analogy in some detail. In particular, I will show that the analogy is to some extent misleading because the 'evidence-based' approach to biomedical research that privileges randomised evaluations is, important as it is, just one approach among others. An alternative, the inferential approach, does not pay any special attention to study design but instead proceeds by eliminating competing hypotheses — by whatever means available in the given case. In what follows, I will describe the inferential approach in some detail and then use it to (a) formulate the debate between 'design-based' and

[1] In fact, virtually all of the philosophical commentary has focused on the virtues and vices of evidence-based approaches in medicine and has, at least so far, ignored social science and policy. The only exceptions I know of are numerous works by Nancy Cartwright (e.g., Cartwright 2009, Cartwright and Munro 2010, Cartwright and Hardie 2012) and the introductory chapter mentioned above (Reiss 2013: Ch. 11) — all of which examine and criticise randomised trials.

‘traditional’ econometrics in its terms, and (b) draw a few lessons from it for this debate. First, however, let us see how the alternative approaches to econometrics solve an important inferential problem.

2. Does Police Reduce Crime? Instruments and Experiments

It is rare that the arrival of a new econometric methodology makes the headlines. Usually, if anything at all, it is new theoretical developments in economics that are of interest to non-specialists: Keynesian economics, monetarism, more recently behavioural economics. But this time is different. The most recent stars of the profession are empirical researchers who run field experiments and manipulate spreadsheets and other computer programmes rather mathematical equations: Joshua Angrist, Esther Duflo, Steven Levitt, to name but a few.

Something else is different. These researchers do not only focus on empirical rather than theoretical analysis — always an important if not exactly prestigious part of economics — but they do so allegedly without much guidance by background theory. To explain, suppose we would like to find out whether some independent variable *I* causes some dependent variable *D*. Observational sciences such as econometrics rely for this matter on the inferences from correlations. As everyone knows, correlation is not causation, and so to find *I* and *D* to be correlated does not go a very long way to establish that *I* causes *D*. Traditionally, a correlational analysis would be carried out against a theoretical background. Consider the standard example in which *I* = quantity supplied for a good = *Q* and *D* = price of that good = *P*. The regression:

(1) $Q = a_S + b_S P + e_S$

is not identified because (as theory informs us) of the existence of a second equation for demand:

(2) $P = a_D + b_D Q + e_D$.

Four parameters cannot be estimated with two equations. However, if background theory also tells us that there is a variable which enters one equation but not the other, for instance that equation (2) is in fact incomplete and should be written thus:

(2′) $P = a_D + b_DQ + c_DZ + e_D$,

then parameter b_S can be identified (not b_D though; to identify b_D we would need a variable that enters (1) but not (2)).

The problem with this standard solution to the standard identification problem is that for most intents and purposes no adequate theories are available that are strong and credible enough to give us econometric equations that allow unambiguous identification of the relevant parameters. The growth literature is a case in point. As Jessica Cohen and William Easterly comment (Cohen and Easterly 2009: 3):

> In the absence of clear evidence that growth outcomes can be attributed to specific levers, development research has severely limited utility for policy. This deficiency was probably due to the infeasibility of instrumenting for multiple RHS variables. Any such attempts usually relied on the Arellano-Bond or Arellano-Bover dynamic panel techniques, which (essentially using lagged RHS variables as instruments) became a kind of magical machine churning out causal econometric results. Unfortunately, the identifying assumptions were so implausible as to leave most outside observers unconvinced. This left the causality question unresolved.

One thing that was new about the recent movement was to turn away from the big questions such as 'What are the causes of growth?' to much smaller questions often concerned with the impact of specific policies or programmes on specific outcomes: 'Does hiring police reduce crime?' (Levitt 1997); 'Which of three school-based HIV/AIDS interventions were most effective in preventing teenage childbearing (as a measure of unprotected sex) in Western Kenya?' (Duflo et al. 2006); 'Does classroom size affect scholastic attainment?' (Angrist and Lavy 1999). The other thing that was new which was already mentioned was that these estimations were made without the benefit of any systematic economic theory. Instead, experimental settings were either created and analysed (Duflo *et al.*) or they occurred naturally and were analysed (Levitt, Angrist and Lavy).

One apparent effect of the experimentalist turn (if we want to call it that) is that the questions these economists address are not only 'smaller' than they used to be, they concern a much broader range of issues than the old, 'big' questions. If economic theory is needed for empirical investigations, presumably, economists can only address economic issues; if, by contrast, all that is needed is a set of workable empirical tools, nothing should stop economists from veering into territory that used to be occupied by social policy researchers, sociologists, educationists, political scientists and others.

Let us look at one of these studies in slightly more detail. Levitt 1997 is interested in measuring the impact of hiring policemen on crime incidence. If we used a regression equation of the form:

(3) $D = a + bI + e$,

where now D = crime incidence (Levitt uses various measures for violent crime, property crime and so on) and I = police (Levitt uses the number of sworn officers), the estimate of b would likely be biased. This because we know that in cities with higher crime rates, more policemen will be hired. Unlike in (1)-(2), however, there is no second equation — instead, institutional background knowledge tells us.

Levitt now employs further institutional facts to solve the problem. He provides evidence that 'Increases in the size of police forces in large cities are disproportionally concentrated in mayoral and gubernatorial election years' (Levitt 1997: 271). He further argues that if the election cycle is to affect the crime rate, then it will be through social spending so that after controlling for spending, election timing will be unrelated to crime. These facts together lead him to believe that the election cycle is a valid instrument.

There are various definitions of instrumental variables (IVs) on offer. It has been argued that there are good reasons to make the causal properties of instruments explicit and therefore to define '*causal* instrumental variables'. A causal instrumental variable Z has three characteristics (Reiss 2005; 2008):

CIV-1: Z causes the independent variable I;
CIV-2: Z affects the dependent variable D, if at all, only through the independent variable I;
CIV-3: Z is not itself caused by the dependent variable D or by a variable that also affects the dependent variable.

These assumptions, together with some general assumptions about the relation between causality and probability and the functional correctness of regression equations can be shown to guarantee instrumental-variable studies to deliver causally correct conclusions (*ibid*.).

That the election cycle is a valid causal instrument is at least plausible. The election cycle is certainly unaffected by crime or by some factor that also affects crime (CIV-3). It is plausible that more police are hired in election years, and Levitt supports this hypothesis with some evidence (CIV-1). Election cycles may affect crime through routes other than those that go through police, but Levitt plausibly argues that this effect will be mediated by public spending, so that after controlling for spending the election timing should be independent of crime (CIV-2).

Randomised studies are very similar, except that the researcher deliberately introduces a 'treatment' (the independent variable) in order to estimate its effect on the 'outcome' (the dependent variable). In fact, it can be shown that randomisation is an instrumental variable (Heckman 1996b). That this should be so is not difficult to see. The randomised assignment causes the treatment/independent variable (CIV-1). Causal routes to the dependent variable other than those going through the independent variable are blocked by the use of a placebo or alternative treatments in the control group, blinding and other techniques (CIV-2). Being random, the assignment is not caused by the dependent variable or shares a common cause with it (CIV-3).

3. Design-Based vs Structural Econometrics

In the eyes of their proponents, the increased use of instrumental variables and randomisation has brought about a 'credibility revolution' in empirical economics (Angrist and Pischke 2010). The idea is that randomised experiments provide a benchmark for empirical work in economics (*ibid*.: 12; original emphasis):

> In applied micro fields such as development, education, environmental economics, health, labor, and public finance, researchers seek real experiments where feasible, and useful natural experiments if real experiments seem (at least for a time) infeasible. In either case, a hallmark of contemporary applied microeconometrics is a conceptual framework that highlights specific

> sources of variation. These studies can be said to be *design based* in that they give the research design underlying any sort of study the attention it would command in a real experiment.

I will follow this terminology and refer to the new methodology as 'design-based econometrics'. Design-based econometrics proceeds by analysing the outcomes of *experimental interventions*. Interventions can either be implemented by the social researcher, as in randomised field experiments, or else 'natural', when an instrumental variable mimics a deliberate intervention. The design of these types of studies matters because not every experimental manipulation or exogenous variable allows causal interpretation. If, say, the intervention is itself caused by a determinant of the outcome variable (for instance because of selection) or the exogenous variable affects the outcome variable through multiple routes, estimates of the causal effect will be biased. Study design aims to eliminate these and other sources of error. Design-based econometrics is atheoretical in that economic theory, according to its proponents, is not required for the identification of a valid experimental intervention or instrumental variable. Key to running an RCT is the successful balancing of treatment and control group with respect to all causal factors relevant to the outcome variable. Causal background knowledge may well play a role in its design but no substantive theory that explains why factors behave in the way they do. Similarly for IV studies — researchers need to know that CIV1-3 are satisfied, but that information can come from causal background knowledge and need not be derived from theory.

The results of design-based econometric studies, then, are said to be credible to the extent that their designs mimic or approach that of an ideal experiment. Does the appeal to experimentation really solve any of the credibility problems of earlier econometrics?

Perhaps not surprisingly, there are influential groups within econometrics who argue that it does not. There are three main critiques. The first is that design-based econometricians search the key where the light is. Instead of asking economically interesting or policy relevant questions, researchers look for natural experiments, valid instruments and treatments whose administering can be randomised independently of whether the question that can be addressed using these designs is worth asking. A corollary of this point is that design-based studies usually identify an *average causal effect* (or related quantities such as the local average treatment effect, LATE), a quantity with limited usefulness (Reiss 2013: Ch. 11).

One problem averages is that they glance over differences in effect size among individuals and subgroups. Reducing class size may have a positive effect on educational attainment overall but ineffective in school *x* and have a negative effect for certain subjects. Any policy will, however, be implemented in individual units for which an average effect, estimated for a population of test units, may be a misleading guide. A local average treatment effect measures the causal effect for those units that were induced by the instrumental variable to take up the treatment — a quantity that is hard to interpret for policy contexts (*cf.* Deaton 2010).

James Heckman, a well-known critic of design-based econometrics distinguishes three policy evaluation problems (Heckman 2008: 7-9):

P1 *Evaluating the Impact of Historical Interventions on Outcomes Including Their Impact in Terms of the Well-Being of the Treated and Society at Large.*

P2 *Forecasting the Impacts (Constructing Counterfactual States) of Interventions Implemented in one Environment in Other Environments, Including Their Impacts In Terms of Well-Being.*

P3 *Forecasting the Impacts of Interventions (Constructing Counterfactual States Associated with Interventions) Never Historically Experienced to Various Environments, Including Their Impacts in Terms of Well-Being.*

The results of typical design-based studies can only address policy problem P1. Policy makers, however, must frequently address problems P2 and P3, for which (according to Heckman) a different methodology is needed.

The second criticism is related. If an IV or randomised study addresses only P1, i.e., it evaluates the impact of a historical intervention, there is no guarantee that the result will also hold elsewhere — in particular in new settings that are of interest to the policy maker. Arguably, studies of *any* kind have this problem. The results of a large-scale cross-country econometric study on growth that uses data from 1960 through 2000 are not guaranteed to hold after 2000 or for countries that are not in the sample or, for that matter, for individual countries *in* the sample. The argument that is sometimes made is that design-based studies, in particular randomised studies, are particularly prone to suffer from this problem.

The reason is that randomisation can introduce artefacts that not likely to persist outside the experimental setting. For example, in a field experiment, experimental subjects may receive more care and attention than outside it. Or subjects might not like the fact that they are part of a randomised experiment — essentially, a lottery — and behave in ways that are uncharacteristic of natural settings in which do not play lotteries (Teira and Reiss 2013). IV analyses of natural experiments do not suffer from these specific problems but it is conceivable that structures in which one can find valid IVs are unrepresentative or uncharacteristic of the policy settings of interest. The main issue in both cases is, however, that the design that is meant to make sure that the study result is correct of the experimental setting in no way speaks in favour of its exportability to the relevant settings.

The third criticism is the most damning one. Structural econometricians and their allies criticise design-based studies by pointing out that instruments, if not backed up by a theoretical or structural model are most likely to be invalid. Angus Deaton, for instance, compares two instances of using IVs for parameter estimation (Deaton 2010). The first starts from a two-equation Keynesian macroeconomic model:

(4) $C = \alpha + \beta Y + u$

(5) $Y \equiv C + I$

It is obvious that investment I is a valid instrument for estimating β in this model. Deaton's second example starts from the following equation:

(6) $P_c = \gamma + \theta R_c + v_c$,

where R_c is a dichotomous variable indicating whether or not city c has a railway station and P_c is a poverty measure. An instrument, following the usual, non-causal definition, is a variable that is correlated with R_c but uncorrelated with v_c. Deaton mentions 'indicators of whether the city has been designated by the Government of China as belonging to a special "infrastructure development area," or perhaps an earthquake that conveniently destroyed a selection of railway stations, or even the existence of river confluence near the city, since rivers were an early source of power, and railways served the power-based industries' as possible examples. Thus we can write:

(7) $R_c = \phi + \varphi Z_c + \eta_c$.

But there is an important difference between (4)&(5) and (6)&(7), respectively, Deaton argues: 'The crucial difference is that the relationship between railways and poverty is not a model at all, unlike the consumption model which embodied a(n admittedly crude) theory of income determination.' Deaton thus seems to regard the existence of a theory-based model as necessary for identifying an instrument. In the second case, there are likely to be many mechanisms that connect the railway stations with poverty, some of which will work in some contexts but not in others. Therefore θ is unlikely to be constant across cities and its variation will not generally be random. (4)&(5) embody, at least in principle, a *theory of income determination*; no such theory of poverty determination is represented by (6)&(7).

There is some ambiguity about what an acceptable model is in this context. Deaton calls the Keynesian macro model 'crude' but he certainly regards it as good enough for his purposes. By contrast, the tradition of structural econometrics that goes back to Haavelmo insists on models describing the behaviour of 'individuals or collective units in their economic activity, their decisions to produce and consume' (Haavelmo 1944: 3), that is, micro models. Whether or not one should insist on micro-based models, what is clear is that purely empirical models such as (6)&(7) won't do.

The problem is, of course, that there are few models like (4)&(5) that enjoy wide acceptance in the economics profession. This is why we needed a 'credibility revolution' in the first place. So we seem to be stuck between a rock and a hard place. On the one hand, we have the neo-empiricists who try to estimate causal parameters without much background theory but whose design-based programme suffers from severe methodological deficits. On the other hand, we have the neo-Cowlesians who think that econometric analyses require the marriage of economic theory, mathematics *and* statistics, and who have at least in-principle answers to their opponents' methodological problems (see in particular Heckman 2008) but whose analyses must build on theoretical background assumptions their opponents are not willing to grant.

4. Two Approaches to Reasoning from Evidence

The design-based approach to empirical economics is sometimes motivated and supported by an appeal to the authority of biomedical science. Here is a quote from Esther Duflo and Michael Kremer (Duflo and Kremer 2005): 'Just as randomized trials revolutionized medicine in the twentieth century, they have the potential to revolutionize social policy during the twenty-first.'

One might counter the Duflo/Kremer defence in a number of possible ways. Any appeal to authority is only as good as the authority appealed to, and perhaps biomedical research does not have the best track record of evidentiary standards. This has certainly been argued before (e.g., by Ioannidis 2005). One might alternatively agree with Duflo and Kremer that randomised trials have revolutionised medicine but not with their (implicit) assessment that this was a revolution for the better.

What I want to do here is draw attention to the fact that in contemporary biomedical research there is not, as Duflo and Kremer suggest, a single 'evidence-based' paradigm that regards randomised trials against the gold standard against which to judge all sources of evidence but rather by two alternative and competing approaches to reasoning from evidence (*cf.* Parascandola 2004). There is indeed the *experimentalist* approach, according to which randomised experiments constitute the 'gold standard' of evidence and all other methods are assessed in terms of how closely they resemble the gold standard. The experimentalist approach is indeed at work in all the domains labeled 'evidence-based', which include parts of medicine, dentistry, nursing, psychology and other fields.

However, there is also the *inferentialist* approach, according to which scientific claims are inferred, using pragmatic criteria, from diverse bodies of evidence that may but need not include experiments. Many scientists across the biomedical sciences subscribe to the inferentialist approach, albeit usually less candidly than the proponents of experimentalism.

The experimental paradigm has received considerable philosophical analysis and support since the times of Bacon and Mill. Indeed, Mill's methods are best understood as accounts of controlled experimentation and more recent work on evidence and causality can be used to underwrite randomised controlled trials (Mayo 1996; Woodward 2003; Cartwright 2007). Even the philosophical literature that takes a critical stance towards evidence-based medicine and practice, tends to focus on the virtues and vices of randomised experimentation.

One reason for the dominance of discussions of the experimental approach also in the methodological literature is that the inferential paradigm is much harder to articulate and defend. The approach seems to raise more questions than it answers: What are the supposed 'pragmatic criteria' on the basis of which hypotheses are to be inferred? What is the body of evidence from which the hypothesis is to be inferred and how diverse does it have to be? If there is no gold standard such as the ideal randomised trial, how do we know what facts to include as evidence?

To present a detailed articulation and defence of the inferentialist approach to reasoning from evidence is beyond the scope of this paper. Instead I will outline one proposal for an articulation of the approach in broad strokes in the remainder of this section and apply these ideas to the debate between design-based and structural econometrics in the next.

The articulation of the inferentialist approach to reasoning from evidence loosely builds on the hypothetico-deductive theory of evidence (Ayer [1936] 1971; Hempel 1966). According to what one might call 'standard hypothetico-deductivism' (standard-HD) a statement *e* is evidence for a hypothesis *h* if and only if *h* deductively entails *e*. Standard-HD is subject to a number of well-known and serious paradoxes and counterexamples (see for instance Hempel 1945).

The account of evidence I defend modifies standard-HD in two main ways. In order to determine what is evidence for a hypothesis instead of demanding that *h* entail *e* it, first, asks 'What patterns in the data would we expect to obtain if *h* were true?'. Thus, taking Levitt's hypothesis that hiring police reduces crime, we would expect police to be negatively correlated with crime under the supposition that the hypothesis is true. The expectation relation is looser than entailment. That *I* and *D* are correlated is something that we'd expect to hold if *I* causes *D* but it is not strictly entailed by the hypothesis. *I* and *D* may fail to be correlated even though *I* causes *D*, for instance, because *I* causes *D* through two routes, one positive and one negative, in such a way that the two routes mutually cancel. Generally speaking, causal hypotheses entail very little about the behaviour of the variables that are causally related. Nevertheless, if the variables are causally related, we would expect the variables to behave in certain ways, and if they happen to behave in these ways we can take this fact as evidence in favour of the causal hypothesis.

The same behaviour of variables can be accounted for in any number of ways. To use the same example, if *I* and *D* are not correlated, then this can be due to a lack of causal connectedness or due to cancelling, due to measurement error and so on. The second required modification of standard-HD is therefore that evidence for a hypothesis consists not only in that which follows from the hypothesis under test (loosely speaking!) but also that which is incompatible with alternative hypotheses. I call the former kind of evidence 'direct', the latter 'indirect'.

In this account of evidence, the context of a scientific inquiry plays a crucial role. Context determines:

(a) what we would expect to obtain under the supposition that the hypothesis is true;
(b) what alternative hypotheses are to be considered relevant;
(c) what it means to rule out or eliminate an alternative hypothesis.

The Empirical Content of a Hypothesis. One feature of the context of a scientific inquiry is given by background knowledge (or background assumptions) about how the world works. In much of the 19th century, the germ theory of disease according to which causes were necessary universal conditions for a disease celebrated significant successes biomedical research. This model eventually led to the adoption of Koch's Postulates:

1. The microorganism must be found in abundance in all organisms suffering from the disease, but should not be found in healthy organisms.
2. The microorganism must be isolated from a diseased organism and grown in pure culture.
3. The cultured microorganism should cause disease when introduced into a healthy organism.
4. The microorganism must be reisolated from the inoculated, diseased experimental host and identified as being identical to the original specific causative agent.

Whether or not causes of diseases behave in the way presupposed by these postulates is an empirical matter. Koch noticed that there were unsymptomatic cholera patients, violating postulates 1 and 3, near the end of the 19th century (Koch 1893). Similarly, causation and determinism were tightly linked until well into the 20th century. Only with the advent of quantum theory became it conceivable that events can be caused and yet not fully determined. Empirical discoveries can affect fundamental beliefs about the nature of causation, domain-specific beliefs about how causes operate and inquiry-specific beliefs about how causes may behave in the narrow context of the inquiry at hand.

Relevant Alternatives. Context also helps to determine which alternatives to a given hypothesis to consider and attempt to eliminate. That an inquiry is made in a scientific context, for instance, proscribes considering sceptical alternatives such as an evil demon hypothesis. Conversely, to consider certain alternatives can change the nature of the debate (*cf.* Williams 2001). If, say, Koch had invoked an evil demon to account for his finding that inoculating some patients with cholera bacteria was not followed by an outbreak of the disease, he would have stopped doing science.

Often epistemic trust requires us to ignore certain alternatives. We could not do science if we suspected deliberate fraud behind every scientific study, for instance. Nevertheless, on occasion we may learn facts about a discipline that make alternatives we would normally ignore relevant. The Vioxx scandal is a case in point (Biddle 2007). The drug was withdrawn from the market in 2004 because of concerns about increased risk of heart attack and stroke. Its manufacturer, Merck, appeared to have known of the risk long before it was withdrawn. In one of the trials in which the drug was tested, five patients in the Vioxx group were reported to have suffered a heart attack or sudden cardiac death, as compared to one in the control group. This difference was not statistically significant. However at least one, and possibly three, more deaths in the Vioxx group resulted from cardiovascular problems but the cause of death was recorded as “unknown”. Had it been classified as a heart attack, the difference in CV events between the Vioxx and naproxen groups would have been statistically significant (Biddle 2007: 29).

What we know about the observation of scientific and ethical standards in a discipline is, thus, a contextual feature that helps to determine which alternatives to regard as relevant. Epistemic trust will normally lead teams of researchers to accept the other teams' results. However, if (say) it is known that death certificates are unreliable when studies are funded by the pharmaceutical industry, a result might and should not be taken at face value but instead probed more deeply.

Eliminating Alternatives. Background knowledge but also normative considerations play important roles in the elimination of alternative hypotheses. In the smoking/lung cancer controversy in the mid-1950s, biomedical researchers were helped by the fact that there is a large dose-response effect between smoking and lung cancer. Moderate smokers increase their risk about ninefold, strong smokers by a factor of 60. This effect is hard to account for by Fisher's constitutional hypothesis, according to which a gene is responsible for the observed association between smoking and the disease. While genes were known to play a role in cancer causation, it was unlikely that genetic factors could not account for the large increase in risk. Individuals with different blood types (which have a genetic basis), for example, differ in cancer susceptibility, but the genetic factor could only account for about 20 percent of the risk (Cornfield et al. 1959).

No amount of evidence can conclusively rule out a relevant alternative. Even if it is highly unlikely that there is a smoking/lung cancer gene that accounts for the dose-response effect, it is of course *possible* that the gene operates in highly unusual ways. Normative considerations help to decide how much evidence that is incompatible with an alternative is needed to consider eliminated. In the smoking/lung cancer case it is probably a lot more harmful to reject the hypothesis if it is in fact true than to accept it if it is false. If smoking does cause lung cancer, people should be warned. Warning people unnecessarily harms the tobacco industry and reduces people's enjoyment of smoking, but it is reasonable to assume that these harms are small as compared to those that would obtain if smoking was responsible for the increase in lung cancer and people would not be warned. Thus, if the consequences of continuing to entertain an alternative hypothesis are relatively benign, a researcher may explain away countervailing evidence but not if they are malignant. At any rate, the elimination of alternatives will always depend on normative considerations of this kind.

That factual and normative considerations are entangled in evidential reasoning is now a familiar point in the philosophy of science (e.g., Douglas 2009, Elliott and McKaughan 2009). The inferentialist approach sketched here makes the issue salient, in at least two ways. First, there does not seem to be a fact of the matter whether or not a body of evidence is incompatible with a given alternative hypothesis. Given the beliefs of the time, genetic factors could clearly account for 20% of the elevated cancer risk among smokers; perhaps it was similarly clear that genetic factors could not account for the 10,000% of risk increase if it existed. There was evidence that for strong smokers the risk increases by about 5,900%. Where to draw the line? All we can say is that with increasing risk the constitutional hypothesis becomes increasingly implausible. But it is one thing to continue to maintain an implausible hypothesis when the cost of doing so is negligible and quite another when it is not, as in the smoking case. Second, any evidence used to eliminate hypothesis is itself just that — evidence. The estimate of a risk elevation of 5,900% is therefore itself uncertain and potentially subject to further scrutiny. Can this figure be accounted for by mismeasurement? Have anti-smoking interest groups financed the studies? Have they been conducted with the appropriate care? Considerations such as these are often crucial, especially in biomedical and social research. Thus, each piece of evidence can itself be probed indefinitely deeply. We obviously have to stop at some point, but there is no fact of the matter where every rational inquirer ought to stop. And again, continuing to maintain that, say, the risk increase for strong smokers is not 5,900% is associated with certain costs and benefits whose relative importance can only be assessed on the basis of normative considerations.

Before moving on to applying the inferentialist approach to econometrics, let me briefly comment on other philosophical work on evidence. Eliminativism is certainly not a novel idea, and it has been defended in particular in the context of evidence-based medicine (e.g., Worrall 2007, Howick 2011). Moreover, other major philosophical theories of evidence such as Bayesianism, error statistics and inference to the best explanation can all be said to be at least consistent with the idea of eliminating alternatives (as was shown for instance in Hawthorne 1993 for Bayesianism, Norton 2005 for error statistics and Bird 2007 for inference to the best explanation).

There is no space here for a full-fledged philosophical defence of the proposed inferentialist theory against all competitors. However, let me briefly say why I believe that an alternative to the standard theories is needed (for a more detailed criticism, see BLINDED). First, Bayesian Confirmation Theory (BCT). According to BCT, a proposition *e* is evidence for a hypothesis *h* just in case P(*h* | *e*) > P(*h*), that is, *e* raises the probability of *h*. BCT is, however, not a good account of the kind of evidential reasoning that is relevant here (by which I mean to bracket the question whether or not it is a good account of *statistical* hypothesis confirmation). First, it has been shown to be neither descriptively or nor normatively adequate of experimental reasoning in biomedical research (Weber 2005: 108ff.). Second, to know whether *e* raises the probability of *h*, a large number of probabilities have to be assumed, probabilities which aren't normally known and can at best be guessed on no good epistemic grounds. One problem is given by the fact that many of the alternative hypotheses are not mutually exclusive — the truth of the causal hypothesis is typically consistent with the existence of common causes, with selection and other biases, with mismeasurement and so on. Third, even if BCT were an adequate account of evidential reasoning in practice and probabilities could reliably be estimated, the Bayesian formalism is at best a framework that can be used to represent the more substantial issues that are going on in reasoning (Earman 1992):

> Moreover, the ability to provide a Bayesian gloss does not mean that Bayesianism has any real explanatory power. Indeed, the eliminative inductivist will see the Bayesian apparatus merely as a tally device to keep track of a more fundamental process.

The inferentialist account sketched here is meant to represent this more fundamental process Earman describes.

The error-statistical theory of evidence maintains that data *d* produced by test *T* is (good) evidence for a hypothesis *h* to the extent that test *T* severely passes *h* with *e*. A severe test is such that the chances that *T* produces similar data are very small if *h* were false (Mayo 1996). Error statistics thus underwrites experimentalism: in an ideal randomised experiment, the chances that the treatment is correlated with the outcome are indeed very low if the treatment did not cause the outcome. Mayo, both in her own work as well as in collaborative work with Aris Spanos, sometimes *says* that the test procedure can consist of a series of tests (e.g., Mayo and Spanos 2010: 25), but the suggestion is never worked out and she never makes clear how different pieces of evidence can fit together to jointly constitute a severe test.

Inference to the best explanation requires the hypothesis not to *entail* the evidence but to *explain* it. Typical scientific hypotheses explain typical evidential statements at best in a very loose sense, however (REF BLINDED). For instance, it is not clear at all whether a causal hypothesis explains a correlation. A causal relation does not cause a correlation nor does is necessarily unify many different phenomena. Typical epistemic virtues such as simplicity, fruitfulness or conservativism do not play a role in the cases of evidential reasoning in the biomedical and social sciences I have looked at here. The proposed account maintains that hypothesis and evidence are inferentially related, and while the inference can sometimes be explanatory, there is no presumption that it has to be (it is therefore *not* one of 'inference to the only explanation' as is Bird 2007).

5. An Inferentialist Perspective on Econometrics

The aim of this section is to use the inferentialist approach to throw some light on arguments given by defenders of design-based and structural econometrics. While the inferentialism has some affinities with structural econometrics, the approach is in fact a lot broader and I will argue that there are defensible ideas behind *both* traditions. I will also give some reasons to believe that some arguments one often hears from either side are ultimately unconvincing.

The inferentialist approach, as described in the previous section, is able to provide rationales for both design-based as well as structural econometrics. The virtues of design-based econometrics lie in the fact that a good design helps to eliminate a number of important alternative hypotheses in one fell swoop. If Levitt's hypothesis that police reduces crime is correct, we should observe a negative correlation between the two variables. Finding the negative correlation, however, can be accounted for in a myriad of different ways. Most significantly, there may be reverse causation: higher crime rates may lead to the general deterioration of an area, which in turn may cause administrators to give up on the area and move police out. In this particular case, a reverse influence that has the opposite sign is more likely: more police will be hired in areas with high crime rates. Nevertheless, a negative reverse influence is possible and should be considered. Moreover, it is possible that a common factor affects both police and crime.

Levitt's design aims to rule out these alternatives. If his instrument is valid and correlated with the crime rate, then there will be a causal route from police to crime (and police reduces crime if the correlation is negative). The same is true of randomised studies. Their design, too, aims to ensure that the outcome cannot cause the treatment and that treatment and outcome cannot have common causes.

The inferentialist approach also motivates why IV studies such as Levitt's are only regarded as 'second best' solution to the estimation problem and randomised trials are more highly regarded. A valid instrument solves the estimation problem just as well as randomisation. In fact, as we have seen above, randomisation simply *is* an instrument. However, when randomisation is achieved through some physical apparatus, it would be very unreasonable to continue to maintain certain alternatives, for instance, that the dependent variable should cause the treatment allocation. IV studies aim to achieve the same, but they rely on more specific causal background knowledge. There will often be contexts in which that can reasonably be challenged. It is at least conceivable, say, that elections are sometimes called in response to changes in the crime rate, invalidating Levitt's choice of instrument.

On the other hand, inferentialism allows for a much broader evidence base than design-based econometrics. Specifically, it does not draw any principled distinction between experimental and observational or randomised and non-randomised studies. Interventions (randomised or not) are but one way to eliminate alternatives. In some contexts they may be an adequate and powerful means for causal inference but there are alternatives.

The structural approach to econometrics can be motivated by the consideration that, in a social science such as economics, it is too easy to come up with plausible relevant alternatives, making a purely empiricist course of action like that promoted by design-based econometrics impossible. Daniel Steel makes this point in the context of using mechanisms for causal inference (Steel 2004: 65):

> The problem lies in the ease of imagining social mechanisms through which nearly any aggregate-level social variable can influence another. Thus, it is rarely the case that no plausible mechanism can be imagined that could connect two variables representing aspects of social phenomena. […]
>
> Listing possible mechanisms through which [for example] opportunity could produce unhappiness does nothing to rule out this plausible alternative. Indeed, this case illustrates how an overabundance of plausible mechanisms is a major source of difficulty for causal inference in the social sciences.

Theory tells us which alternatives to consider and which are irrelevant because not motivated by theory. James Heckman, for example, argued that the random sequence number (which determined whether a member of a cohort of males would be drafted to fight in the Vietnam war) used by Angrist as an instrument to estimate the effect of serving in the military on civil earnings was invalid because in a behavioural model managers would observe their employees' random sequence number, determine the probability that an employee would be drafted and make investment in training decisions dependent on this probability (Heckman 1996a: 461). Theory thus tells us that there is a route from instrument to dependent variable that does not go through the independent variable, violating CIV-2.

The inferentialist framework outlined here can also show where proponents of either side are mistaken. Design-based econometricians are mistaken to assume that only good study design is able to convincingly rule out alternatives. Good design is neither necessary nor sufficient to eliminate *all* alternatives. Levitt's study shows both. Given that the influence of crime on police is overwhelmingly likely to be positive, there is in fact no need to use an instrument to argue that police reduces crime, if the variables are negatively correlated and if establishing a purely qualitative causal claim is the aim of the inquiry. Context-specific background knowledge and taking into account of the purpose of the inquiry suffice. (The argument so far would leave the possibility of a common cause. However, here too we may have context-specific reasons for setting aside alternatives according to which common factors affect police in one direction and crime in the other.)

Moreover, good design serves only to eliminate some relevant alternatives, never all of them. For instance, John McCrary tried to replicate Levitt's study and found that a weighting error in Levitt's computer programme led to the estimate that hiring police is effective in reducing crime, which, when the error is corrected, disappears (McCrary 2002). Using a valid instrument or randomisation does not protect against mismeasurement, sloppiness, fraud and other sources of error. To give a couple of further examples, when David Neumark and William Wascher tried to replicate David Card and Alan Krueger's famous New Jersey-Pennsylvania minimum wage experiment (and IV study) using payroll data instead of telephone surveys, they come to the opposite conclusion — namely, that the increase in the minimum wage led to a *de*crease in employment (Card and Krueger 1994; Neumark and

Wascher 2000). A 1977 analysis of four negative income tax field experiments in the United States finds that providing individuals with a guaranteed income destabilises families and increases the divorce rate; a 1990 re-analysis (which, among other things, pools the data differently) finds no such thing (Hannan et al. 1977; Cain and Wissoker 1990).

These sources of error have to be ruled out by means other than design. Peer review and replication may help to eliminate coding and programming errors. Calibration, careful investigation of the measurement procedure and sensitivity analyses may protect against measurement error. Design in the narrow, experimentalist sense alone won't do. The other side of this coin is that if it is realised that evidence that has nothing to do with the design of the study is needed to rule out certain alternatives anyway, it will appear more palatable to use these other kinds of evidence to are rule out the causal alternatives such as reverse causation/ simultaneity and common causes as well.

Something similar can be said about structural econometrics. I have no qualms with the idea that economic theory can help make alternatives relevant and eliminate them. It would be preposterous to maintain, however, that theory is necessary for the job or sufficient or both. It may well be the case that transient overdetermination is a serious issue in the social sciences because it is so easy to come up with plausible hypotheses. But economics is inundated by data as well and there is no guarantee that there are no contexts in which all relevant alternatives can be eliminated on the basis of data alone.

Conversely, theory is clearly not sufficient to make hypotheses relevant or eliminate them either. Economic theory would probably make us expect a lot more fraud than there actually is (because it makes us believe that people do what they can get away with). Even if there is a plausible and widely accepted theoretical model within which some variable is a valid IV, the model may of course just be false. What matters is that the variable has the requisite causal properties CIV-1-3, not whether we learned that it does from a model, from institutional or other background knowledge or from other sources.

6. Conclusions: Judgements All Over the Place

The inferentialist theory of evidence outlined here emphasises the role of *judgements* in inferences from evidence to hypothesis. We require judgements to know what patterns in the data to expect if a hypothesis were true; judgements to know what alternative hypotheses to consider relevant (and what patterns in the data to expect if any of them were true); and judgements to know at what point an alternative should be considered eliminated on the basis of data that is incompatible with it.

Judgement smacks of subjectivity or, worse, arbitrariness, and many scientists and methodologists prefer to replace it by strict rules. Design-based and structural econometricians agree on this: while one camp uses the rule 'Mimic a controlled experiment as closely as possible!' and the other the rule 'Deduce econometric models from theory!', both prefer rules to judgement.

There is no getting around judgement in inferring a hypothesis from the evidence, however. Even in the ideal, totally controlled experiment judgements have to be made about the relevance of causal factors for the outcome of interest. No two situations are ever exactly alike. They can at best be alike with respect to relevant causal factors. In IV studies we have to judge whether the purported instrument is a valid CIV, in randomised experiments whether treatment and control group are balanced with respect to causal factors. In the structural approach to econometrics, when empirical analyses are conducted against theoretical models, we rely on our judgement to determine whether the model is a good one (e.g., Does it include all relevant variables? Is it functionally correct? Does it include lags of appropriate lengths?).

In the eyes of many, the role of judgement in inference is like the role of the government to the neo-liberal. Unlike their libertarian brethren, neo-liberals allow the government an important place in the social order. As James Buchanan, in his 1986 presidential address to the Mont Pèlerin Society, explains (Buchanan 1986: 2):

> For most of our members, however, social order without a state is not readily imagined, at least not in any normatively preferred sense... Man is, and must remain, a slave to the state. But it is critically and vitally important to recognize that ten percent slavery is different from fifty percent slavery.

Substitute 'judgement' for 'state', 'causal inference' for 'social order' and 'econometrician' for 'man', and we have a statement with which most econometricians, design-based or structural, would agree, I believe:

> For most of our members, however, causal inference without judgement is not readily imagined, at least not in any normatively preferred sense... The econometrician is, and must remain, a slave to judgement. But it is critically and vitally important to recognize that ten percent slavery is different from fifty percent slavery.

The correct response (in both cases) is of course that not the amount counts but the quality. We need to get our judgements *right*, quite independently of *how much* judgement there is in inference.

Acknowledgements

I wish to thank audiences in London, Durham, Madrid and Rotterdam, where previous versions of this paper were given, for valuable comments and discussion, as well as the editors of this special issue and two anonymous referees. This work was supported by the Spanish Ministry of Science and Innovation under Grant FFI2012-37354/CONSOLIDER INGENIO CSD2009-0056.

Bibliography

Angrist, J. and J.-S. Pischke (2010). The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics, National Bureau of Economic Research. 15794.

Angrist, J. D. and V. Lavy (1999). "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement." The Quarterly Journal of Economics 114(2): 533-575.

Ayer, A. ([1936] 1971). Language, Truth and Logic. London, Penguin Books.

Biddle, J. (2007). "Lessons from the Vioxx Debacle: What the Privatization of Science Can Teach Us About Social Epistemology." Social Epistemology 21(1): 21-39.

Bird, A. (2007). "Inference to the Only Explanation." Philosophy and Phenomenological Research 74(2): 424-432.

Buchanan, J. (1986). Man and the State. Mont Pèlerin Society Presidential Address. M. P. Society, Liberaal Achief, Ghent.

Cain, C. G. and D. A. Wissoker (1990). "A Reanalysis of Marital Stability in the Seattle-Denver Income-Maintenance Experiment." American Journal of Sociology 95(5): 235-314.

Card, D. and A. Krueger (1994). "Minimum Wages and Employment: A Case Study of the Fast Food Industry in New Jersey and Pennsylvania." American Economic Review 84(4): 772-793.

Cartwright, N. (2007). "Are RCTs the Gold Standard?" BioSocieties 2(2): 11-20.

Cartwright, N. (2009). "Evidence-Based Policy: What's to Be Done About Relevance." Philosophical Studies 143(1): 127-136.

Cartwright, N. and J. Hardie (2012). Evidence-Based Policy: A Practical Guide to Doing it Better. Oxford, Oxford University Press.

Cartwright, N. and E. Munro (2010). "The limitations of randomized controlled trials in predicting effectiveness." Journal of Evaluation in Clinical Practice 16(2): 260-266.

Cohen, J. and W. Easterly (2009). Introduction: Thinking Big versus Thinking Small. What Works in Development. Thinking Big and Thinking Small. Washington, DC, Brookings Institution: 1-23.

Cornfield, J., W. Haenszel, C. Hammond, A. Lilienfield, M. Shimkin and E. Wynder (1959). "Smoking and lung cancer: recent evidence and a discussion of some questions." Journal of the National Cancer Institute 22: 173-203.

Deaton, A. (2010). "Instruments, randomization, and learning about development." Journal of Economic Literature 48(2): 424-455.

Douglas, H. (2009). Science, Policy, and the Value-Free Ideal. Pittsburgh, University of Pittsburgh Press.

Duflo, E., P. Dupas, M. Kremer and S. Sinei (2006). Education and Hiv/Aids Prevention: Evidence from a Randomized Evaluation in Western Kenya. Policy Research Working Paper. Washington (DC), World Bank. 4024.

Duflo, E. and M. Kremer (2005). Use of Randomization in the Evaluation of Development Effectiveness. Evaluating Development Effectiveness. G. Pitman, O. Feinstein and G. Ingram. New Brunswick (NJ), Transaction Publishers. 7.

Earman, J. (1992). Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory. Cambridge (MA), MIT Press.

Elliott, K. and D. McKaughan (2009). "How Values in Scientific Discovery and Pursuit Alter Theory Appraisal." Philosophy of Science 76(PSA 2008 Proceedings): 598-611.

Haavelmo, T. (1944). "The Probability Approach in Econometrics." Econometrica 12(Supplement): iii-115.

Hannan, M., N. Tuma and L. Groeneveld (1977). "Income and marital events: evidence from an income-maintenance experiment." American Journal of Sociology 82: 1186-1211.

Hawthorne, J. (1993). "Bayesian Induction is Eliminative Induction." Philosophical Topics 21(1): 99-138.

Heckman, J. (1992). Randomization and Social Policy Evaluation. Evaluating Welfare and Training Programs. C. F. Manski and I. Garfinkel. Boston (MA), Harvard University Press: 201-230.

Heckman, J. (1996). "Randomization as an Instrumental Variable." The Review of Economics and Statistics 78(2): 336-341.

Heckman, J. (1996a). "Comment." Journal of the American Statistical Association 91(434): 459-462.

Heckman, J. (2008). Econometric Causality. Discussion Paper. Bonn, IZA.

Hempel, C. (1945). "Studies in the Logic of Confirmation (I.)." Mind 54(213): 1-26.

Hempel, C. (1966). The Philosophy of Natural Science. Upper Saddle River (NJ), Prentice-Hall.

Howick, J. (2011). The Philosophy of Evidence-Based Medicine. Chichester, Wiley-Blackwell.

Ioannidis, J. (2005). "Why Most Published Research Findings Are False." PLoS Medicine 2(8): e124.

Koch, R. (1893). "Über den augenblicklichen Stand der bakteriologischen Choleradiagnose." Zeitschrift für Hygiene und Infectionskrankheiten 14: 319-333.

Levitt, S. (1997). "Using Electoral Cycles in Police Hiring to Estimate the Effect of Police on Crime." American Economic Review 87(3): 270-290.

Mayo, D. (1996). Error and the Growth of Experimental Knowledge. Chicago, University of Chicago Press.

Mayo, D. and A. Spanos (2010). Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Rationality of Science. Cambridge, Cambridge University Press.

McCrary, J. (2002). "Using Electoral Cycles in Police Hiring to Estimate the Effect of Police on Crime: Comment." American Economic Review 92(4): 1236-1243.

Neumark, D. and W. Wascher (2000). "The Effect of New Jersey's Minimum Wage Increase on Fast-Food Employment: A Reevaluation Using Payroll Records." American Economic Review 90(5): 1362-1396.

Norton, J. (2005). A Little Survey of Induction. Scientific Evidence: Philosophical Theories and Applications. P. Achinstein. Baltimore (MD), The Johns Hopkins University Press: 9-34.

Parascandola, M. (2004). "Two approaches to etiology: the debate over smoking and lung cancer in the 1950s." Endeavour 28(2): 81-86.

Reiss, J. (2005). "Causal Instrumental Variables and Interventions." Philosophy of Science 72(PSA 2004).

Reiss, J. (2008). Error in Economics: Towards a More Evidence-Based Methodology. London, Routledge.

Reiss, J. (2013). Philosophy of Economics: A Contemporary Introduction. New York (NY), Routledge.

Scriven, M. (2008). "A Summative Evaluation of RCT Methodology & An Alternative Approach to Causal Research." Journal of MultiDisciplinary Evaluation 5(9): 11-24.

Steel, D. (2004). "Social Mechanisms and Causal Inference." Philosophy of the Social Sciences 34(1): 55-78.

Teira, D. and J. Reiss (2013). Causality, Impartiality and Evidence-Based Policy. Towards the Methodological Turn in the Philosophy of Science: Mechanism and Causality in Biology and Economics. H.-K. Chao, S.-T. Chen and R. Millstein. New York (NY), Springer: 207-224.

Weber, M. (2005). Philosophy of Experimental Biology. Cambridge, Cambridge University Press.

Williams, M. (2001). Problems of Knowledge. Oxford, Oxford University Press.

Woodward, J. (2003). Making Things Happen. Oxford, Oxford University Press.

Worrall, J. (2007). "Evidence in Medicine and Evidence-Based Medicine." Philosophy Compass 2: 981-1022.