KEYNES LECTURE IN ECONOMICS

Instruments of Development: Randomisation in the Tropics, and the Search for the Elusive Keys to Economic Development

ANGUS DEATON
Fellow of the Academy

1. Introduction

THE EFFECTIVENESS OF DEVELOPMENT ASSISTANCE is a topic of great public interest. Much of the public debate among non-economists takes it for granted that, if the funds were made available, development would follow: Pogge (2005), Singer (2004), and at least some economists agree (Sachs 2005, 2008). Others, most notably Easterly (2006), are deeply sceptical, a position that has been forcefully argued at least since Bauer (1971, 1981). Few academic economists or political scientists agree with Sachs' views, but there is a wide range of intermediate positions, recently well assembled by Easterly (2008). The debate runs the gamut from the macro—can foreign assistance raise growth rates and eliminate poverty?—to the micro—what sorts of projects are likely to be effective?; should aid focus on electricity and roads, or on the provision of schools and clinics or vaccination campaigns? In this lecture, I shall be concerned with both the macro and micro kinds of assistance. I shall have very little to say about what actually works and what does not; but it is clear from the literature that we do not know. Instead, my main concern is with how we should go

Read at the Academy 9 October 2008.

Proceedings of the British Academy, **162**, 123–160. © The British Academy 2009.

about finding out whether and how assistance works and with methods for gathering evidence and learning from it in a scientific and cumulative way. I am not an econometrician, but I believe that econometric methodology needs to be assessed, not only by methodologists but also by those who are concerned with the substance of the issue. Only they (we) are in a position to tell when something has gone wrong with the application of econometric methods, not because they are incorrect given their assumptions, but because their assumptions do not apply, or because they are incorrectly conceived for the problem at hand. Or at least that is my excuse for meddling in these matters.

Any analysis of the extent to which foreign aid has increased economic growth in recipient countries immediately confronts the familiar problem of simultaneous causality; the effect of aid on growth, if any, will be disguised by effects running in the opposite direction, from poor economic performance to compensatory or humanitarian aid. It is not immediately obvious how to disentangle these effects, and some have argued that the question is not answerable and that econometric studies of it should be abandoned. Certainly, the econometric studies that use international evidence to examine aid effectiveness currently have low professional status. Yet it cannot be right to give up on the issue. There is no general or public understanding that nothing can be said, and to give up the econometric analysis is simply to abandon precise statements for loose and unconstrained histories of episodes selected to support the position of the speaker.

The analysis of aid effectiveness typically uses cross-country growth regressions with the simultaneity between aid and growth dealt with using instrumental variable methods. I shall argue in the next section that there has been a good deal of misunderstanding in the literature about the use of instrumental variables. Econometric analysis has changed its focus over the years, away from the analysis of models derived from theory towards much looser specifications that are statistical representations of programme evaluation. With this shift, instrumental variables have moved from being solutions to a well-defined problem of inference to being devices that induce quasi-randomisation. Old and new understandings of instruments co-exist, leading to errors, misunderstandings and confusion, as well as unfortunate and unnecessary rhetorical barriers between disciplines working on the same problems. These abuses of technique have contributed to a general scepticism about the ability of econometric analysis to answer these big questions.

A similar state of affairs exists in the microeconomic area, in the analysis of the effectiveness of individual programmes and projects, such as the construction of infrastructure—dams, roads, water supply, electricity—and in the delivery of services—for example for education, health or policing. There is great frustration with aid organisations, particularly the World Bank, for allegedly failing to learn from its projects and to build up a systematic catalogue of what works and what does not. In addition, some of the scepticism about macroeconometrics extends to microeconometrics, so that there has been a movement away from such methods and towards randomised controlled trials. According to Esther Duflo, one of the leaders of the new movement in development, 'Creating a culture in which rigorous randomized evaluations are promoted, encouraged, and financed has the potential to revolutionize social policy during the 21st century, just as randomized trials revolutionized medicine during the 20th.' This quote is from a 2004 *Lancet* editorial headed 'The World Bank is finally embracing science'.

In Section 4 of this paper, I shall argue that in ideal circumstances randomised evaluations of projects are useful for obtaining a convincing estimate of the average effect of a programme or project. The price for this success is a focus that is too narrow to tell us 'what works' in development, to design policy, or to advance scientific knowledge about development processes. Project evaluation using randomised controlled trials is unlikely to discover the elusive keys to development, nor to be the basis for a cumulative research programme that might progressively lead to a better understanding of development. This argument applies *a fortiori* to instrumental variables strategies that are aimed at generating quasi-experiments; the value of econometric methods cannot be assessed by how closely they approximate randomised controlled trials. Following Cartwright (2007*a*, 2007*b*), I argue that evidence from randomised controlled trials has no special priority. Randomisation is not a gold standard because 'there is no gold standard' (Cartwright 2007*a*). Randomised controlled trials cannot automatically trump other evidence, they do not occupy any special place in some hierarchy of evidence; nor does it make sense to refer to them as 'hard' while other methods are 'soft'. These rhetorical devices are just that; a metaphor is not an argument.

More positively, I shall argue that the analysis of projects needs to be refocused towards the investigation of potentially generalisable mechanisms that explain why and in what contexts projects can be expected to work. The best of the experimental work in development economics

already does so, because its practitioners are too talented to be bound by their own methodological prescriptions. Yet there would be much to be said for doing so more openly. I concur with the general message in Pawson and Tilley (1997), who argue that thirty years of project evaluation in sociology, education and criminology was largely unsuccessful because it focused on *whether* projects work instead of on *why* they work. In economics, warnings along the same lines have been repeatedly given by James Heckman—see particularly Heckman (1992) and Heckman and Smith (1995)— and much of what I have to say is a recapitulation of his arguments.

The paper is organised as follows. Section 2 lays out some econometric preliminaries concerning instrumental variables and the vexed question of exogeneity. Section 3 is about aid and growth. Section 4 is about randomised controlled trials. Section 5 is about using empirical evidence and where we should go next.

2. Instruments, identification, and the meaning of exogeneity

It is useful to begin with a simple and familiar econometric model that I can use to illustrate the differences between different flavours of econometric practice; this has nothing to do with economic development, but has the virtue of simplicity and is easy to contrast with the development practice that I wish to discuss. In contrast to the models that I will discuss later, I think of this as a model in the spirit of the Cowles Foundation. It is the simplest possible Keynesian macroeconomic model of national income determination taken from once-standard econometrics textbooks. There are two equations which together comprise a complete macroeconomic system. The first equation is a consumption function, in which aggregate consumption is a linear function of aggregate national income, while the second is the national income accounting identity which says that income is the sum of consumption and investment. I write the system in standard notation as

$$C = \alpha + \beta Y + u \tag{1}$$

$$Y = C + I \tag{2}$$

According to (1), consumers choose the level of aggregate consumption (C) with reference to their income (Y), while in (2) investment is set by the 'animal spirits' of entrepreneurs in a way that is outside of the model. No

modern macroeconomist would take this model seriously, though the simple consumption function is clearly an ancestor of more satisfactory and complete modern formulations; in particular, we can think of it (or at least its descendants) as being derived from a coherent model of intertemporal choice. Similarly, modern versions would postulate some theory for what determines investment I; here it is simply taken as given, and assumed to be orthogonal to the consumption disturbance u.

In this model, consumption and income are simultaneously determined so that, in particular, a stochastic realisation of u—consumers displaying animal spirits of their own—will affect not only C but also Y through equation (2), so that there is a positive correlation between u and Y. As a result, ordinary least squares estimation of (1) will lead to upwardly biased and inconsistent estimates of the parameter β.

This simultaneity problem can be dealt with in a number of ways. One is to solve (1) and (2), to get the reduced form equations

$$C = \frac{\alpha}{1-\beta} + \frac{\beta}{1-\beta} I + \frac{u}{1-\beta} \tag{3}$$

$$Y = \frac{\alpha}{1-\beta} + \frac{I}{1-\beta} + \frac{u}{1-\beta} \tag{4}$$

Either of these equations can be consistently estimated by OLS, and it is easy to show that the same estimates of α and β will be obtained from either one. An alternative method of estimation is to focus on the consumption function (1), and to use our knowledge of (2) to note that investment can be used as an instrumental variable (IV) for income. In the IV regression, there is a 'first stage' regression in which income is regressed on investment; this is identical to equation (4), which is part of the reduced form. In the second stage, consumption is regressed on the predicted value of income from (4). In this simple case, the IV estimate of β is identical to the estimate from the reduced form. This simple model may not be a very good model, but it is a model, if only a primitive one.

I now leap forward sixty years, and consider an apparently similar set up, again using an absurdly simple specification. The World Bank (let us imagine) is interested in whether to advise the government of China to build more railway stations. The Bank economists write down an econometric model in which the poverty head count ratio in city c is taken to be a linear function of an indicator R of whether or not the city has a railway station,

$$P_c = \gamma + \theta R_c + v_c \tag{5}$$

where θ (I hesitate to call it a parameter) indicates the effect—presumably negative—of infrastructure (here a railway station) on poverty. While we cannot expect to get useful estimates of θ from OLS estimation of (5)—railway stations may be built to serve more prosperous cities, they are rarely built in deserts where there are no people, or there may be 'third factors' that influence both—this is seen as a 'technical problem' for which there is a wide range of econometric treatments including, of course, instrumental variables.

We no longer have the reduced form of the previous model to guide us but if we can find an instrument Z that is correlated with whether a town has a railway station, but uncorrelated with v, we can do the same calculations and obtain a consistent estimate. For the record, I write this equation

$$R_c = \varnothing + \varphi Z_c + \eta_c \tag{6}$$

Good candidates for Z might be 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 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. I am making fun, but not much.

My main argument is that the two econometric structures, in spite of their resemblance and the fact that IV techniques can be used for both, are in fact quite different. In particular, the IV procedures that work for the effect of national income on consumption are unlikely to give useful results for the effect of railway stations on poverty. To explain the differences, I begin with the language. In the original example, consumption and income are treated symmetrically, and appear as such in the reduced form equations (3) and (4). In contemporary examples, such as the railways, there is no symmetry. Instead, we have a 'main' equation (5), which used to be the 'structural' equation (1). We also have a 'first-stage' equation, which is the regression of railway stations on the instrument, which was previously just one of the equations in the reduced form. The reduced form, of course, was typically more completely specified than the first-stage regression, since it was derived from a notionally complete model of the system. The now rarely considered regression of the variable of interest on the instrument, here of poverty on earthquakes or on river confluences, is nowadays referred to as *the* reduced form, although it was originally one equation of a multiple equation reduced form within which

it had no special significance. These language shifts sometimes cause confusion, but they are not the most important differences between the two systems.

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. While it is clearly *possible* that the construction of a railway station will reduce poverty, there are many possible mechanisms, some of which will work in one context and not in another. In consequence θ is unlikely to be constant over different cities, nor can its variation be usefully thought of as random variation that is uncorrelated with anything else of interest. Instead, it is precisely the variation in θ that encapsulates the poverty reduction mechanisms that ought to be the main objects of our enquiry. Instead, the equation of interest is thought of as a representation of something more akin to an experiment or a biomedical trial, in which some cities get 'treated' with a station, and some do not. The role of econometric analysis is not, as in the Cowles example, to estimate and investigate a casual model, but 'to create an analogy, perhaps forced, between an observational study and an experiment' (Freedman 2006: 691).

One immediate task is to recognise and somehow deal with the variation in θ, which is typically referred to as the 'heterogeneity problem' in the literature. The obvious way is to define a parameter of interest in a way that corresponds to something we want to know for policy evaluation—perhaps the average effect on poverty over some group of cities—and then devise an appropriate estimation strategy. However, this step is often skipped in practice, perhaps because of a mistaken belief that (5) is a structural equation in which θ is a constant, so that the analysis can go immediately to the choice of instrument Z, over which a great deal of imagination is often exercised. As in the traditional model, the instrument is selected to satisfy two criteria, that it be correlated with R_c and uncorrelated with v_c. Of course, if heterogeneity is indeed present the probability limit of the IV estimator will in general depend on the choice of instrument (Heckman 1997). Such a procedure is the opposite of standard statistical practice, in which a parameter of interest is defined first, followed by an estimator that delivers that parameter. Instead, we have a procedure in which the choice of the instrument, which is guided by criteria designed for a different situation, is implicitly allowed to determine the parameter of interest. This goes beyond the old story of looking at an object where the light is strong enough to see; rather, we have control over

the light, but choose to let it fall where it may, and then proclaim that whatever it illuminates is what we were looking for all along.

Recent econometric analysis has given us a more precise characterisation of what we can expect from such a method. In the railway example, where the instrument is the designation of a city as belonging to the 'special infrastructure zone', the probability limit of the IV estimator is the average of poverty reduction effects over those cities that were induced to construct a railway station by being so designated. This average is known as the 'local average treatment effect' (LATE), and its recovery by IV estimation requires a number of non-trivial conditions including, for example, that no cities where a railway station would have been constructed are perverse enough to be actually deterred from doing so by the positive designation (see Angrist and Imbens 1994, who established the LATE theorem). The LATE may, or may not, be a parameter of interest to the World Bank or the Chinese government and in general there is no reason to suppose that it will be. For example, the parameter estimated will typically *not* be the average poverty reduction effect over the designated cities, nor the average effect over all cities.

I find it hard to make any sense of the LATE. We are unlikely to learn much about the processes at work if we refuse to say *anything* about what determines θ; heterogeneity is not a technical problem calling for an econometric solution, but is a reflection of the fact that we have not started on our proper business, which is trying to understand what is going on. Of course, if we are as sceptical of the ability of economic theory to deliver useful models as are many applied economists today, the ability to avoid modelling can be seen as an advantage, though it should not be a surprise when such an approach provides answers that are hard to interpret.

There is a related issue that bedevils a good deal of contemporary applied work, which is the understanding of *exogeneity*, a word that I have so far avoided. Suppose, for the moment, that the effect of railway stations on poverty is the same in all cities, and we are looking for an instrument which is required to be exogenous in order to consistently estimate θ. According to Merriam-Webster's dictionary, 'exogenous' means 'caused by factors or an agent from outside the organism or system', and this common usage is often employed in applied work. However, the consistency of IV estimation requires that the instrument be orthogonal to the error term v in the equation of interest, which is not implied by the Merriam-Webster definition (see Leamer 1985: 260). Heckman (2000) suggests using the term 'external' (which he traces back to Wright and

Frisch in the 1930s) for the Merriam-Webster definition, for variables whose values are not set or caused by the variables in the model (according to this, consumption and investment are 'internal' variables, and investment an 'external' variable), and keeping 'exogenous' for the orthogonality condition that is required for consistent estimation in this instrumental variable context. The terms are hardly standard, but I adopt them here because I need to make the distinction. The main issue, however, is not the terminology but that the two concepts be kept distinct, so that we can see when the argument being *offered* is a justification for externality when what is *required* is exogeneity. An instrument that is external, but not exogeneous, will not yield consistent estimates of the parameter of interest, even when that parameter is constant.

Failure to separate externality and endogeneity has caused, and continues to cause, endless confusion in the applied development (and other) literatures. Natural or geographic variables—distance from the equator (as an instrument for per capita GDP in explaining religiosity; McCleary and Barro 2006), rivers (as an instrument for the number of school districts in explaining educational outcomes; Hoxby 2000), land gradient (as an instrument for dam construction in explaining poverty; Duflo and Pande 2007), month of birth (as an instrument for years of schooling in an earnings regression; Angrist and Krueger 1991), or rainfall (as an instrument for economic growth in explaining civil war; Miguel, Satyanath, and Sergenti 2004)—the examples could be multiplied *ad infinitum*—are not affected by the variables being explained, and are clearly external. So are historical variables—the mortality of colonial settlers is not influenced by current institutional arrangements in ex-colonial countries (Acemoglu, Johnson and Robinson 2001), nor does the country's growth rate today influence which country they were colonised by (Barro 1998). Whether any of these instruments is *exogenous* depends on the nature of the equation of interest, and is not guaranteed by its *externality*. And because exogeneity is an identifying assumption that must be made prior to analysis of the data, no empirical tests are possible. This does not prevent many attempts in the literature, often by misinterpreting a satisfactory *over*identification test as evidence for valid identification. Such tests can tell us whether estimates change when we select different subsets from a set of possible instruments. While the test is clearly informative, acceptance is consistent with all of the instruments being invalid, while failure is consistent with a subset being correct.

In my running example, earthquakes and rivers are external to the system, and are not caused by either poverty or by the construction of railway

stations, and the designation as an infrastructure zone may also be determined by factors independent of poverty or railways. But even earthquakes (or rivers) are not exogenous if they have an effect on poverty other than through their destruction (or encouragement) of railway stations, as will almost always be the case. The absence of simultaneity does not guarantee exogeneity; exogeneity requires the absence of simultaneity, but is not implied by it. Even random numbers—the ultimate external variables—may be endogenous, at least in the presence of heterogeneity. Again, the example comes from Heckman's (1997) discussion of Angrist's (1990) famous use of draft lottery numbers as an instrumental variable in his analysis of the subsequent earnings of Vietnam veterans.

I can illustrate Heckman's argument using the Chinese railways example with the zone designation as instrument. Rewrite the equation of interest (5) as

$$P_c = \gamma + \bar{\theta} R_c + w_c = \gamma + \bar{\theta} R_c + \{v_c + (\theta - \bar{\theta}) R_c\} \tag{7}$$

where w_c is defined by the term within $\{\}$, and $\bar{\theta}$ is the mean of θ over the cities that get the station so that the compound error term w has mean zero. Suppose the designation as an infrastructure zone is D_c, which takes values 1 or 0, and that the Chinese bureaucracy, persuaded by young development economists, decides to randomise and sets the designation of cities by flipping a *Yuan*. For consistent estimation of $\bar{\theta}$, we want the covariance of the instrument with the error to be zero. The covariance is

$$E(D_c w_c) = E[(\theta - \bar{\theta}) RD] = E[(\theta - \bar{\theta}) | D = 1, R = 1] P(D = 1, R = 1) \tag{8}$$

which will be zero if either (*a*) the average effect of building a railway station on poverty among the cities induced to build one by the designation is the same as the average effect among those who would have built one anyway, or (*b*) no city not designated builds a railway station.[1] If (*b*) is not guaranteed by *fiat*, we cannot suppose that it will otherwise hold, and we might reasonably hope that among the cities that build railway stations, those induced to do so by the designation are those where there is the largest effect on poverty, which violates (*a*). In the example of the Vietnam veterans, the instrument (the draft lottery number) fails to be exogenous because the error term in the earnings equation depends on each individual's rate of return to schooling, and whether or not each potential draftee accepted their assignment—their veteran's status—depends on that rate of return. In practice, most instruments are not random num-

[1] I am grateful to Winston Lin for clarification on this point.

bers, and the assumption that the instrument is orthogonal to v_c, which is accepted in (8), will also have to be defended.

The general lesson here is once again the ultimate futility of trying to avoid thinking about how and why things work; if we do not do so, we are left with undifferentiated heterogeneity that is likely to prevent consistent estimation of any parameter of interest. One appropriate response is to specify exactly how cities respond to their designation, an approach that leads to Heckman's local instrumental variable methods (Heckman and Vytlacil 1999, 2007; Heckman, Urzua and Vytlacil 2006). Similar questions are pursued by van den Berg (2008).

3. Instruments of development

The question of whether aid has helped economies grow faster is typically asked within the framework of standard growth regressions. These regressions use data for many countries over a period of years, usually from the Penn World Table, the current version of which provides data on real per capita GDP and its components in purchasing power dollars for more than 150 countries as far back as 1950. The model to be estimated has the rate of growth of per capita GDP as the dependent variable, while the explanatory variables include the lagged value of GDP per capita, the share of investment in GDP, and measures of the educational level of the population (see, for example, Barro and Sala-i-Martin 1995, chapter 12, for an overview). Other variables are often added, and my main concern here is with one of these, external assistance (aid) as a fraction of GDP. A typical specification can be written

$$\Delta 1\mathrm{n}\, Y_{ct+1} = \beta_0 + \beta_1 1\mathrm{n}\, Y_{ct} + \beta_2 \frac{I_{ct}}{Y_{ct}} + \beta_3 H_{ct} + \beta_4 Z_{ct} + \theta A_{ct} + u_{ct} \qquad (9)$$

where Y is per capita GDP, I is investment, H is a measure of human capital or education, and A is the variable of interest, aid as a share of GDP. Z stands for whatever other variables are included. The index c is for country and t for time. Growth is rarely measured on a year to year basis—the data in the Penn World Table are not suitable for annual analysis—so that growth may be measured over ten-, twenty-, or forty-year intervals. With around forty years of data, there are four, two, or one observation for each country.

An immediate question is whether the growth equation (9) is a model-based Cowles-type equation, as in my national income example, or

whether it is more akin to the atheoretical analyses in my invented Chinese railway example. There are elements of both here. If we ignore the Z and A variables in (9), the model can be thought of as a Solow growth model, extended to add human capital to physical capital—see again Barro and Sala-i-Martin (1995), who derive their empirical specifications from the theory, and also Mankiw, Romer and Weil (1992), who extended the Solow model to include education. However, the addition of the other variables, including aid, is typically less well justified. In some cases, for example under the assumption that all aid is invested, it is possible to calculate what effect we might expect aid to have (see Rajan and Subramanian 2005). If we follow this route, (9) would not be useful—because aid is already included—and we should instead investigate *whether* aid is indeed invested, and then infer the effectiveness of aid from the effectiveness of investment. Even so, it presumably matters what kind of investment is promoted by aid, and aid for roads, for dams, for vaccination programmes, or for humanitarian purposes after an earthquake is likely to have different effects on subsequent growth. More broadly, one of the main issues of contention in the whole debate is what aid actually does. Just to list a few of the possibilities, does aid increase investment, does aid crowd out domestic investment, is aid stolen, or does aid create rent-seeking that undercuts the long-run conditions for growth? Once all of these possibilities are admitted, it is clear that the analysis of (9) is not a Cowles model at all, but is seen as some sort of biomedical experiment in which different countries are 'dosed' with different amounts of aid, and we are trying to measure the average response. As in the Chinese railways case, a regression such as (9) will not give us what we want, because the doses of aid are not randomly administered to different countries, so our first task is to find an instrumental variable that will generate quasi-randomness.

The most obvious problem with a regression of aid on growth is the simultaneous feedback from growth to aid that is generated by humanitarian responses to either economic collapse or natural or human-made disasters that engender economic collapse. More generally, aid flows from rich countries to poor countries, and poor countries, by definition, are those with poor records of economic growth. This feedback, from low growth to high aid, will obscure, nullify, or reverse any positive effects of aid. Most of the literature attempts to eliminate this feedback by using one or more instrumental variables and, although they would not express it in these terms, the aim of the instrumentation is to a restore a situation in which the pure effect of aid on growth can be observed, as if in a

randomised situation. How close we get to this ideal depends, of course, on the choice of instrument.

Although there is some variation across studies, there is a standard set of instruments, originally proposed by Boone (1996), which include the log of population size and various country dummies, for example, a dummy for Egypt, or for francophone West Africa. One or both of these instruments are used in almost all the papers in a large subsequent literature, including Burnside and Dollar (2000), Hansen and Tarp (2000, 2001), Dalgaard and Hansen (2001), Guillamont and Chauvet (2001), Lensink and White (2001), Easterly, Levine, and Roodman (2003), Dalgaard, Hansen, and Tarp (2004), Clemens, Radelet, and Bhavani (2004), Rajan and Subramanian (2005), and Roodman (2008). The rationale for population size is that larger countries get less aid per capita, because the aid agencies allocate aid on a country basis, with less than full allowance for population size. The rationale for what I shall refer to as the 'Egypt instrument' is that Egypt gets a great deal of American aid as part of the Camp David accords in which it agreed to a partial rapprochement with Israel. The same argument applied to the francophone countries, which receive additional aid from France because of their past colonial status. By comparing these countries with countries not so favoured, or by comparing populous with less populous countries, we can observe a kind of variation in the share of aid in GDP that is unaffected by the negative feedback from poor growth to compensatory aid. In effect, we are using the variation across populations of different sizes as a natural experiment to reveal the effects of aid.

If we examine the effects of aid on growth without any allowance for reverse causality, for example by estimating equation (9) by ordinary least squares, the estimated effect is typically negative. For example, Rajan and Subramanian (2005), in one of the most careful recent studies, find that an increase in aid by 1 per cent of GDP comes with a reduction in the growth rate of one tenth of a percentage point a year. Easterly (2005) paints with a broader brush, and provides many other (sometimes spectacular) examples of negative associations between aid and growth. When instrumental variables are used to eliminate the reverse causality, Rajan and Subramanian find a weak or zero effect of aid, and contrast that finding with the robust positive effects of investment on growth in specifications like (9). I should note that although Rajan and Subramanian's study is an excellent one, it is certainly not without its critics and, as the authors note, there are many difficult econometric problems beyond the choice of instruments, including how to estimate dynamic models with country

fixed effects on limited data, the choice of countries and sample period, the type of aid that needs to be considered, and so on. Indeed, it is those other issues that are the focus of most of the literature cited above. The substance of this debate is far from over.

My main concern here is with the use of the instruments, what they tell us, and what they might tell us. The first point is that neither the 'Egypt' nor the population instrument are plausibly exogenous; both are external—Camp David is not part of the model, nor was it caused by Egypt's economic growth, and similarly for population size—but exogeneity would require that neither 'Egypt' nor population size have any influence on economic growth except through the effects on aid flows, which makes no sense at all. We also need to recognise the heterogeneity in the aid responses, and try to think about how the different instruments are implicitly choosing different averages, involving different weightings of countries. Or we could stop right here, conclude that there are no valid instruments, and that the aid to growth question is not answerable in this way. I shall argue otherwise, but I should also note that similar challenges over the validity of instruments have become routine in applied econometrics, leading to widespread scepticism by some, while others press on undaunted in an ever more creative search for exogeneity.

Yet consideration of the instruments is not without value, especially if we move away from instrumental variable estimation, with the use of instruments seen as technical, not substantive, and think about the reduced form which contains substantive information about the relationship between growth and the instruments. For the case of population size, we find that, conditional on the other variables, population size is unrelated to growth, which is one of the reasons that the IV estimates of the effects of aid are small or zero. This (partial) regression coefficient is a much simpler object than is the instrumental variable estimate; under standard assumptions, it tells us how much faster large countries grow than small countries, once the standard effects of the augmented Solow model have been taken into account. Does this tell us anything about the effectiveness of aid? Not directly, though it is surely useful to know that although large countries receive less per capita aid in relation to per capita income, they have grown just as fast as countries that have received more, once we take into account the amount that they invest, their levels of education, and their starting level of GDP. But we would hardly conclude from this fact alone that aid does not increase growth. Perhaps aid works less well in small countries, or perhaps there is an offsetting positive effect of population size on economic growth. Both are possible, and

both are worth further investigation. More generally, such arguments are susceptible to fruitful discussions, not only among economists but also with other social scientists and historians who study these questions, something that is typically difficult with instrumental variables. Economists' claims to methodological superiority based on instrumental variables ring particularly hollow when it is economists themselves who are often misled. My argument is that for both economists and non-economists, the direct consideration of the reduced form is likely to generate productive lines of enquiry.

The case of the 'Egypt' instrument is somewhat different. Once again the reduced form is useful (Egypt doesn't grow particularly fast in spite of all the aid it gets after Camp David), though mostly for making it immediately clear that the comparison of Egypt versus non-Egypt, or francophone versus non-francophone, is not a useful way of assessing the effectiveness of aid on growth. Yet almost every paper in this literature unquestioningly uses the Egypt dummy as an instrument.

I conclude this section with an example that helps bridge the gap between analyses of the macro and analyses of the micro effects of aid. Many microeconomists agree that instrumentation in cross-country regressions is unlikely to be useful, while claiming that microeconomic analysis is capable of doing better. We may not be able to answer ill-posed questions about the macroeconomic effects of foreign assistance, but we can surely do better on specific projects and programmes. Banerjee and He (2008) have provided a list of the sort of studies that they like and that they believe should be replicated more widely. One of these, also endorsed by Duflo (2004), is a famous paper by Angrist and Lavy (1999) on whether schoolchildren do better in smaller classes, a position frequently endorsed by parents and teachers' unions, but not always supported by empirical work. The question is an important one for development assistance, because smaller class sizes cost money, and are a potential use for foreign aid. Angrist and Lavy's paper uses a natural experiment, not a real one, and relies on IV estimation, so it provides a bridge between the relatively weak natural experiments in this section, and the actual randomised controlled trials in the next.

Angrist and Lavy's study is about the allocation of children enrolled in a school into classes. Many countries set their class sizes to conform to some version of Maimonides' Rule, which sets a maximum class size, beyond which additional teachers must be found. In Israel, the maximum class size is set at 40. If there are less than 40 children enrolled, they will all be in the same class. If there are 41, there will be two classes, one of

20, and one of 21. If there are 81 or more children, the first two classes will be full, and more must be set up. Angrist and Lavy's Figure 1 plots actual class size and Maimonides' Rule class size against the number of children enrolled; this graph starts off running along the forty-five-degree line, and then falls discontinuously to 20 when enrolment is 40, increasing with slope of 0.5 to 80, falling to 27.7 (80 divided by 3) at 80, rising again with a slope of 0.25, and so on. They show that actual class-sizes, while not exactly conforming to the rule, are strongly influenced by it, and exhibit the same saw-tooth pattern. They then plot test scores against enrolment, and show that they display the opposite pattern, rising at each of the discontinuities where class-size abruptly falls. This is a natural experiment, with Maimonides' Rule inducing quasi-experimental variation, and generating a predicted class size for each level of enrolment which serves as an instrumental variable in a regression of test scores on class size. These IV estimates, unlike the OLS estimates, show that children in smaller class sizes do better.

Angrist and Lavy's paper, the creativity of its method, and the clarity and convincingness of its result has set the standard for micro-empirical work since it was published, and it has had a far-reaching effect on subsequent empirical work in labour and development economics. Yet there is a problem, which has become apparent over time. Note first the heterogeneity; it is improbable that the effect of lower class size is the same for all children so that, under the assumptions of the LATE theorem, the IV estimate recovers a weighted average of the effects for those children who are shifted by Maimonides' Rule from a larger to a smaller class. Those children might not be the same as other children, which makes it hard to know how useful the numbers might be in other contexts, for example when all children are put in smaller class sizes. The underlying reasons for this heterogeneity are not addressed in this quasi-experimental approach. To be sure of what is happening here, we need to know more about how different children finish up in different classes, which raises the possibility that the variation across the discontinuities may not be orthogonal to other factors that affect test scores.

A recent paper by Urquiola and Verhoogen (2009) explores how it is that children are allocated to different class sizes in a related, but different, situation in Chile where a version of Maimonides' Rule is in place. Urquiola and Verhoogen note that parents care a great deal about whether their children are in the 40 child class or the 20 child class and, for the private schools they study, they construct a model in which there is sorting across the boundary, so that the children in the smaller classes

have richer, more educated parents than the children in the larger classes. Their data match such a model, so that at least some of the differences in test scores across class size come from differences in the children that would be present whatever the class-size. This paper is an elegant example of why it is so dangerous to make inferences from natural experiments without understanding the mechanisms at work.

In preparation for the next section, I note that the problem here is *not* the fact that we have a quasi-experiment rather than a real experiment, so that there was no actual randomisation. If children had been randomised into classes of varying size, the problems would have been the same, unless there had been some mechanism for forcing the children (and their parents) to accept the assignment.

4. Randomisation in the tropics

Scepticism about econometrics, doubts about the usefulness of structural models in economics, and the endless wrangling over identification and instrumental variables, have led to a search for alternative ways of learning about development. There has also been frustration with the World Bank's apparent failure to learn from its own projects, and its inability to provide a convincing argument that its past activities have enhanced economic growth and poverty reduction. Past development practice is seen as a succession of fads, with one supposed magic bullet replacing another—from planning to infrastructure to human capital to structural adjustment to health and social capital to the environment and back to infrastructure—a process that seems not to be guided by progressive learning. For many economists, and particularly for the group at the Poverty Action Lab at MIT, the solution has been to move towards randomised controlled trials (RCTs) of projects, programmes and policies. RCTs are seen as generating gold standard evidence that is superior to econometric evidence, and that is immune to the methodological criticisms that have been characteristic of econometric analyses. Another aim of the programme is to persuade the World Bank to replace its current evaluation methods with RCTs; Duflo (2004) argues that randomised trials of projects would generate knowledge that could be used elsewhere, an international public good. Banerjee (2007*a*, chapter 1) accuses the Bank of 'lazy thinking', of a 'resistance to knowledge', and notes that its recommendations for poverty reduction and empowerment show a striking 'lack of distinction made between strategies founded on the hard evidence

provided by randomized trials or natural experiments and the rest'. In all this there is a close parallel with the evidence-based movement in medicine that preceded it, and the successes of RCTs in medicine are frequently cited. Yet the parallels are almost entirely rhetorical, and there is little or no reference to the dissenting literature, as surveyed for example by Worrall (2007), who documents the rise and fall in medicine of the rhetoric used by Banerjee. Nor is there any recognition of the many problems of medical RCTs.

The movement in favour of RCTs is currently very successful. The World Bank is now conducting substantial numbers of randomised trials, and the methodology is sometimes explicitly requested by governments, who supply the World Bank with funds for this purpose (see World Bank 2008 for details of the Spanish Trust Fund for Impact Evaluation). There is a new International Initiative for Impact Evaluation which 'seeks to improve the lives of poor people in low- and middle-income countries by providing, and summarizing, evidence of what works, when, why and for how much' (International Initiative for Impact Evaluation 2008). The Poverty Action Lab lists dozens of completed and ongoing projects in a large number of countries, many of which are project evaluations. Many development economists would subscribe to the jingoist view proclaimed by the editors of the *British Medical Journal* (quoted by Worrall 2007) which noted that 'Britain has given the world Shakespeare, Newtonian physics, the theory of evolution, parliamentary democracy—and the randomized trial'.

4.1 The ideal RCT

Under ideal conditions, and when correctly executed, an RCT can estimate certain quantities of interest with minimal assumptions, thus absolving RCTs of one complaint against econometric methods, that they rest on often implausible economic models. It is useful to lay out briefly the (standard) framework for these results, originally due to Jerzy Neyman in the 1920s, currently often referred to as the Holland–Rubin framework or the Rubin causal model (see Freedman 2006, for a discussion of the history). According to this, each member of the population under study, labeled i, has two possible values associated with it, Y_{0i} and Y_{1i}, which are the outcomes that i would display if it did not get the treatment, $T_i = 0$ and if it did get the treatment, $T_i = 1$. Since each i is either in the treatment group or in the control group, we observe one of Y_{0i} and Y_{1i}, but not both. We would like to know something about the distribu-

tion over i of the effects of the treatment, $Y_{i1}-Y_{i0}$, in particular its mean $\bar{Y}_1-\bar{Y}_0$. In a sense, the most surprising thing about this set-up is that we can say anything useful at all, without further assumptions, or without any modelling. But that is the magic that is wrought by the randomisation.

What we *can* observe in the data is the difference between the average outcome in the treatments and the average outcome in the controls, or $E(Y_i|T_i=1)-E(Y_i|T_i=0)$. This difference can be broken up into two terms

$$\begin{aligned} E(Y_{i1}|T_i=1)-E(Y_{i0}|T_i=0)=&[E(Y_{i1}|T_i=1)-E(Y_{i0}|T_i=1)] \\ &+[E(Y_{i0}|T_i=1)-E(Y_{i0}|T_i=0)] \end{aligned} \tag{10}$$

Note that on the right-hand side the second term in the first square bracket cancels out with the first term in the second square bracket. But the term in the second square bracket is zero by randomisation; the non-treatment outcomes, like any other characteristic, are identical in expectation in the control and treatment groups. We can therefore write (10) as

$$E(Y_{i1}|T_i=1)-E(Y_{i0}|T_i=0)=[E(Y_{i1}|T_i=1)-E(Y_{i0}|T_i=1)] \tag{11}$$

so that the difference in the two observable outcomes is the difference between the average treated outcome and the average untreated outcome in the treatment group. The last term on the right-hand side would be unobservable in the absence of randomisation.

We are not quite done. What we would like is the average of the difference, rather than the difference of averages that is currently on the right-hand side of (11). But the expectation is a linear operator, so that the difference of the averages is identical to the average of the differences, so that we reach, finally

$$E(Y_{i1}|T_i=1)-E(Y_{i0}|T_i=0)=E(Y_{i1}-Y_{i0}|T_i=1) \tag{12}$$

The difference in means between the treatments and controls is an estimate of the average treatment effect among the treated which, since the treatment and controls differ only by randomisation, is an estimate of the average treatment effect for all. This standard but remarkable result depends both on randomisation and on the linearity of expectations.

One immediate consequence of this derivation is a fact that is often quoted by critics of RCTs but is often ignored by practitioners, at least in economics: RCTs are informative about the *mean* of the treatment effects, $Y_{i1}-Y_{i0}$, but do not identify other features of the distribution. For example, the median of the difference is not the difference in medians, so an RCT is not, by itself, informative about the median treatment effect, something that could be of as much interest to policy makers as the mean

treatment effect. It might also be useful to know the fraction of the population for which the treatment effect is positive, which once again is not identified from a trial. Put differently, the trial might reveal an average positive effect although nearly all of the population is hurt with a few receiving very large benefits, a situation that cannot be revealed by the RCT although it might be disastrous if implemented. Indeed, Kanbur (2001) has argued that much of the disagreement about development policy is driven by differences of this kind. Given the minimal assumptions that go into an RCT, it is not surprising that it cannot tell us everything that we would like to know. Heckman and Smith (1995) discuss these issues at greater length, and also note that, in some circumstances, more can be learned. Essentially, the RCT gives us two marginal distributions, from which we would like to infer a joint distribution; this is impossible, but the marginal distributions limit the joint distribution in a way that can be useful, for example if the distribution among the treated stochastically dominates the distribution among the controls.

In practice, researchers who conduct randomised controlled trials often present results on statistics other than the mean. For example, the results can be used to run a regression of the form

$$Y_i = \beta_0 + \beta_1 T_i + \sum_j \theta_j X_{ij} + \sum_j \phi_j X_{ij} \times T_i + u_i \tag{13}$$

where T is a binary variable that indicates treatment status, and the Xs are various characteristics measured at baseline that are included in the regression both on their own (main effects) and as interactions with treatment status (see de Mel, McKenzie, and Woodruff 2008 for an example of a field experiment with micro-enterprises in Sri Lanka). The estimated treatment effect now varies across the population, so that it is possible, for example, to estimate whether the average treatment effect is positive or negative for various subgroups of interest. These estimates depend on more assumptions than the trial itself, in particular on the validity of running a regression like (13), on which I shall have more to say below. One immediate charge against such a procedure is data mining. A sufficiently determined examination of any trial will eventually reveal some subgroup for whom the treatment yielded a significant effect of some sort, and there is no general way of adjusting standard errors to protect against the possibility. In drug trials, the FDA rules require that analytical plans be submitted prior to trial, and at least one economic experiment—moving to opportunity—has imposed similar rules on itself (see the protocol by Feins and McInnis 2001).

I am not arguing against post-trial subgroup analysis, only that any special epistemic status (as in 'gold standard', 'hard', or 'rigorous' evidence) possessed by RCTs does not extend to subgroup analysis if only because there is no general guarantee that a new RCT on post-experimentally defined subgroups will yield the same result. Yet such analyses do not share any special evidential status that might be accorded to RCTs, and must be assessed in exactly the same way as we would assess any non-experimental or econometric study. These issues are wonderfully exposed by the subgroup analysis of drug effectiveness by Horwitz *et al.* (1996), criticised by Altmann (1998), who refers to such studies as 'a false trail', Senn and Harrell (1997), 'wisdom after the event', and by Davey Smith and Egger (1998), 'incommunicable knowledge', drawing the response 'reaching the tunnel at the end of the light', by Horwitz *et al.* (1997). It is clearly absurd to discard data because we do not know how to analyse them with sufficient purity. Indeed, many important findings have come from post-trial analysis of experimental data, both in medicine and in economics, for example of the negative income tax experiments of the 1960s. None of which resolves the concerns about data-mining. In large-scale, expensive, trials, a zero or very small result is unlikely to be welcome, and there is likely to be overwhelming pressure to search for some subpopulation that gives a more palatable result.

The mean treatment effect from an RCT may be of limited value for a physician or a policymaker contemplating specific patients or policies. A new drug might do better than a placebo in an RCT, yet a physician might be entirely correct in not prescribing it for a patient whose characteristics, according to the physician's theory of the disease, might lead her to suppose that the drug would be harmful. Similarly, if we find that dams in India do not reduce poverty on average, as in Duflo and Pande's fine (non-experimental) 2007 study, there is no implication about any specific dam, even one of the dams included in the study, yet it is always a specific dam that a policy maker has to decide about. Their evidence certainly puts a higher burden of proof on those proposing a new dam, as would be the case for a physician prescribing in the face of an RCT, though the force of the evidence depends on the size of the mean effect and the extent of the heterogeneity in the responses. As was the case with the material discussed in Sections 2 and 3, heterogeneity poses problems for the analysis of RCTs, just as it posed problems for non-experimental methods that sought to approximate randomisation. For this reason, in his *Planning of Experiments* Cox (1958: 15) begins his book with the *assumption* that the treatment effects are identical. He notes that the RCT will still estimate

the mean treatment effect with heterogeneity, but argues that such estimates are 'quite misleading', citing the example of two different subgroups within which the treatment effects are identical so that the RCT delivers an estimate that applies to no one. This recommendation makes a good deal of sense when the experiment is being applied to the parameter of a well-specified model, but it could not be further away from most current practice in either medicine or economics.

One of the reasons why subgroup analysis is so hard to resist is that researchers, however much they may wish to escape the straitjacket of theory, inevitably have some mechanism in mind, and some of those mechanisms can be 'tested' on the data from the trial. Such 'testing', of course, does not satisfy the strict evidential standards that the RCT has been set up to satisfy, and if the investigation is constrained to satisfy those standards no *ex post* speculation is permitted. Without a prior theory, and within its own evidentiary standards, an RCT targeted at 'finding out what works' is not informative about mechanisms. For example, when two independent but identical RCTs in two cities in India find that children's scores improved less in Mumbai than in in Vadodora, the authors state 'this is likely related to the fact that over 80 per cent of the children in Mumbai had already mastered the basic language skills the program was covering' (Duflo, Glennerster, and Kremer 2008). It is not clear how 'likely' is established here, and there is certainly no evidence that conforms to the 'gold standard' that is seen as one of the central justifications for the RCTs. For the same reason, repeated *successful* replications of a 'what works' experiment is both unlikely and unlikely to be persuasive. Learning about theory, or mechanisms, requires that the investigation be targeted towards that theory, towards *why* something works, not *whether* it works. Projects can rarely be replicated, though the mechanisms underlying success or failure will often be replicable and transportable. This means that if the World Bank had indeed randomised all of its past projects, it is unlikely that the cumulated evidence would contain the key to economic development.

Cartwright (2007*a*) summarises the benefits of RCTs relative to other forms of evidence. In the ideal case, 'if the assumptions of the test are met, a positive result *implies* the appropriate causal conclusion' that the intervention 'worked' and caused a positive outcome. She adds that 'the benefit that the conclusions follow deductively in the ideal case comes with great cost: narrowness of scope' (p. 11).

4.2 Tropical RCTs in practice

How well do actual RCTs approximate the ideal? Are the assumptions generally met in practice? Is the narrowness of scope a price that brings real benefits, or is the superiority of RCTs largely rhetorical? As always, there is no substitute for examining each study in detail, and there is certainly nothing in the RCT methodology itself that grants immunity from problems of implementation. Yet a number of general points are worth discussion.

The first is the seemingly obvious practical matter of how to compute the results of a trial. In theory this is straightforward, we simply compare the mean outcome in the experimental group with the mean outcome in the control group, and the difference is the causal effect of the intervention. This simplicity, compared with the often baroque complexity of econometric estimators, is seen as one of the great advantages of RCTs, both in generating convincing results and in explaining those results to policy makers and the lay public. Yet any difference is not useful without a standard error, and the calculation of the standard error is rarely quite so straightforward. As Fisher (1935) emphasised from the very beginning, in his famous discussion of the tea lady, randomisation plays two separate roles. The first is to guarantee that the probability law governing the selection of the control group is the same as the probability law governing the selection of the experimental group. The second is to provide a probability law that enables us to judge whether a difference between the two groups is significant. In his tea lady example, Fisher uses combinatoric analysis to calculate the exact probabilities of each possible outcome, but in practice this is rarely done.

Duflo, Glennerster and Kremer (2008: 3921 (DGK)) explicitly recommend what seems to have become the standard method in the development literature, which is to run a restricted version of the regression (13), including only the constant and the treatment dummy,

$$Y_i = \beta_0 + \beta_1 T_i + u_i \tag{14}$$

As is easily shown, the OLS estimate of β_1 is simply the difference between the mean of the in the experimental and control groups, which is exactly what we want. However, the *standard error* from the OLS regression is not correct unless the variance in the experimental group is identical to the variance in the control group, which will only be true if the treatment has no effect on the variance, which will not generally be the

case particularly if treatment responses are heterogeneous. (Indeed, assuming that the experiment *does not* affect the variance is very much against the minimalist spirit of RCTs.) If the regression (14) is run with the standard heteroscedasticity correction to the standard error, the result will be the same as the formula for the standard error of the difference between two means, but not otherwise unless there are equal numbers of experimental treatments and controls, in which case the correction makes no difference, and the OLS standard error is correct. It is not clear in the experimental development literature whether the correction is routinely done in practice, and the handbook review by DGK makes no mention of it, although it provides a thoroughly useful review of many other aspects of standard errors.

Even with the correction for unequal variances, we are not quite done. The general problem of testing the significance of the differences between the means of two normal populations with different variances is known as the Fisher-Behrens problem. The test statistic computed by dividing the difference in means by its estimated standard error does not have the t-distribution when the variances are different, and the significance of the estimated difference in means is likely to be overstated if no correction is made. If there are equal numbers of treatments and controls, the statistic will be approximately distributed as Student's t, but with degrees of freedom that can be as little as half the nominal degrees of freedom when one of the two variances is zero. In general, there is also no reason to suppose that the heterogeneity in the treatment effects is normal, which will further complicate inference in small samples.

Another standard practice, recommended by DGK, and which is also common in medical RCTs according to Freedman (2008), is to run the regression (14) with additional controls taken from the baseline data, or equivalently (13) with the X_i but without the interactions,

$$Y_i = \beta_0 + \beta_1 T_i + \sum_j \theta_j X_{ij} + u_i \tag{15}$$

The standard argument is that, if the randomisation is done correctly, the X_i are orthogonal to the treatment variable T_i so that their inclusion does not affect the estimate of β_1, which is the parameter of interest. However, by absorbing variance, as compared with (14), they will increase the precision of the estimate—this is not necessarily the case, but will often be true. DGK (p. 3924) give an example that 'controlling for baseline test scores in evaluations of educational interventions greatly improves the

precision of the estimates, which reduces the cost of these evaluations when a baseline test can be conducted'.

There are two problems with this procedure. The first, which is noted by DGK, is that, as with post-trial subgroup analysis, there is a risk of data mining—trying different control variables until the experiment 'works'—unless the control variables are specified in advance. Again, it is hard to tell whether or how often this dictum is observed. The second problem is analysed by Freedman (2008), who notes that (15) is not a standard regression because of the heterogeneity of the responses. Write for the (hypothetical) treatment response of unit i, so that, in line with the discussion in the previous subsection, $a_i = Y_{i1} - Y_{i0}$, and we can write the identity

$$Y_i = Y_{i0} + a_i T_i = \bar{Y}_0 + a_i T_i + (Y_{i0} - \bar{Y}_0) \tag{16}$$

which looks like the regression (15) with the Xs and the error term capturing the variation in Y_{i0}. The only difference is that the coefficient on the treatment term has an i suffix because of the heterogeneity. If we define $\alpha = E(a_i | T_i = 1)$, the average treatment effect among the treated, as the parameter of interest, as in Section 4.1, we can rewrite (16) as

$$Y_i = \bar{Y}_0 + \alpha T_i + (Y_{i0} - \bar{Y}_0) + (a_i - \alpha) T_i \tag{17}$$

Finally, and to illustrate, suppose that we model the variation in Y_{i0} as a linear function of an observable scalar X_{i0} and a residual η_i, we have

$$Y_i = \beta_0 + \alpha T_i + \theta(X_i - \bar{X}) + [\eta_i + (a_i - \alpha) T_i] \tag{18}$$

with $\beta_0 = Y_0$, which is in the regression form (15), but allows us to see the links with the experimental quantities.

Equation (18) is analysed in some detail by Freedman (2008). It is easily shown that T_i is orthogonal to the compound error, but that this is not true of $X_i - \bar{X}$. (Nor will it be true of the interaction terms in equation 15.) However, the two right-hand-side variables are uncorrelated because of the randomisation, so the OLS estimate of β_1 is consistent. This is not true of θ, though this may not be a problem if the aim is to reduce the sampling variance. A more serious issue is that the dependency between T_i and the compound error term means that the OLS estimate is biased, and in small samples this bias—which comes from the heterogeneity—may be substantial. Freedman notes that, in the case where, without loss of generality, X has unit variance, the leading term in the bias of the estimate of the OLS estimate of α is φ/n where n is the sample size and

$$\varphi = -\lim \frac{1}{n} \sum_{i=1}^{n} (a_i - a)(X_i - \bar{X})^2 \tag{19}$$

Equation (19) shows that the bias comes from the heterogeneity or, more specifically, from a covariance between the heterogeneity in the treatment effects and the squares of the included covariates. With the sample sizes typically encountered in these experiments, which are often expensive to conduct, the bias can be substantial. One possible strategy here would be to compare the estimates of a with and without covariates; even ignoring pre-test bias it is not clear how to make such a comparison without a good estimate of the standard error. Another possibility is to introduce the interactions between the covariates and the treatment, which leads back to (13). An incomplete analysis suggests that this reduces the bias compared with (19).

Of these and related issues in medical trials, Freedman writes 'Practitioners will doubtless be heard to object that they know all this perfectly well. Perhaps, but then why do they so often fit models without discussing assumptions?'

All of the issues so far can be dealt with, either by appropriately calculating standard errors or by refraining from the use of covariates, though this might involve drawing larger and more expensive samples. However, there are other practical problems that are harder to fix. One of these is that subjects may fail to accept assignment, so that people who are assigned to the experimental group may refuse, and controls may find a way of getting the treatment, and either may drop out of the experiment altogether. The classical remedy of double blinding, so that neither the subject nor the experimenter know which subject is in which group, is rarely feasible in social experiments—children know their class size—and is often not feasible in medical trials—subjects may decipher the randomisation, for example by asking a laboratory to check that their medicine is not a placebo. Heckman (1992) notes that, in contrast to people, 'plots of grounds do not respond to anticipated treatments of fertilizer, nor can they excuse themselves from being treated'. This makes the important point, further developed by Heckman in later work, that the deviations from assignment are almost certainly purposeful, at least in part. The people who struggle to escape their assignment will do so more vigorously the higher are the stakes, so that the deviations from assignment cannot be treated as random measurement error, but will compromise the results in fundamental ways.

Once again, there is a widely used technical fix, which is to run regressions like (15) or (18), with *actual* treatment status in place of the *assigned* treatment status T_i. This replacement will destroy the orthogonality between treatment and the error term, so that OLS estimation will no longer yield a consistent estimate of the average treatment effect among the treated. However, the assigned treatment status, which is known to the experimenter, is orthogonal to the error term, and is correlated with the actual treatment status, and so can serve as an instrumental variable for the latter. But now we have come all the way back to the discussion of instrumental variables in Section 2, and we are doing econometrics, not an ideal RCT. Under the assumption of no 'defiers'—people who do the opposite of their assignment just because of the assignment (and it is not clear 'just why are there no defiers'; Freedman 2006)—the instrumental variable converges to the local average treatment effect (LATE). As before, it is unclear whether this is what we want, and there is no way to find out without modelling the behaviour that is responsible for the heterogeneity of the response to assignment, as in the local instrumental variable approach developed by Heckman and Vytlacil (1999, 2007). Alternatively, and as recommended by Freedman (2004: 4, 2006), it is always informative to make a simple unadjusted comparison of the average outcomes between treatments and controls according to the original assignment. This may also be enough if what we are concerned with is whether the treatment works or not, rather than with the size of the effect. In terms of instrumental variables, this is a recommendation to look at the reduced form, and again harks back to similar arguments in Section 2 on aid effectiveness.

There is also a host of operational problems that afflict every actual experiment; these can be mitigated by careful planning—in RCTs, compared with econometric analysis, most of the work is done before data collection not after—but not always eliminated. In this context, I turn to the flagship study of the new movement in development economics, Miguel and Kremer's (2004) study of intestinal worms in Kenya. This paper is repeatedly cited in DGK's manual, and it is one of the exemplary studies cited by Duflo (2004) and Banerjee and He (2008). It was written by two senior authors at leading research universities, and published in the most prestigious technical journal in economics. It has also received a great deal of positive attention in the popular press (see, for example, Leonhardt 2008), and has been influential in policy (see Poverty Action Lab 2007). In this study, a group of 'seventy-five rural primary schools

were phased into treatment in a randomized order', with the finding 'that the program reduced school absenteeism by at least one quarter, with particularly large participation gains among the youngest children, making deworming a highly effective way to boost school participation among young children' (p. 159). The point of the RCT is less to show that deworming medicines are effective but rather, because children infect one another, that school-based treatment is more effective than individual treatment. As befits a paper that aims to change methods as well as practice, there is emphasis on the virtues of randomisation, and the word 'random' or its derivatives appears some sixty times in the paper. But the actual method of randomisation is not precisely described in the paper, and private communication with Michael Kremer has confirmed that, in fact, the local partners would not permit the use of random numbers for assignment, so that the assignment of schools to three groups was done in alphabetical order, as in Albert to group 1, Alfred to group 2, Bob to group 3, Charles to group 1 again, David to group 2, and so on. Alphabetisation, not randomisation, was also used in the experiment on flip charts in schools by Glewwe, Kremer, Moulin, and Zitzewitz (2004); this paper, like 'Worms', is much cited as evidence in favour of the virtues of randomisation.

Alphabetisation may be a reasonable solution when randomisation is impossible, but we are then in the world of quasi- or natural experiments, not randomised experiments. As is true with all forms of quasi-randomisation, alphabetisation does not guarantee orthogonality with potential confounders. Resources are often allocated alphabetically, because that is how many lists are presented, and it is easy to imagine that the Kenyan government or local NGOs, like Miguel and Kremer, used the alphabetical list to prioritise projects or funding. If so, schools higher in the alphabet are systematically different, and this difference will be inherited in an attenuated form by the three groups. Indeed, this sort of contamination is described by Cox (1958: 74–5), who explicitly warns against this sort of convenient design. Of course, it is also possible that the alphabetisation causes no confounding with factors known or unknown. If so, there is still an issue with the calculation of standard errors. Without a probability law, we have no way of discovering whether the difference between treatments and controls could have risen by chance. We might think of modelling the situation here by imagining that the assignment was equivalent to taking a random starting value and assigning every third school to treatment. If so, the fact that there are only three possible assignments of schools would have to be taken into account in calculating the standard errors, and nothing of this kind is reported. As it is, it is

impossible to tell whether the experimental differences in these studies are or are not due to chance.

In this subsection I have dwelt on practice not to critique particular studies or particular results; indeed it seems entirely plausible that deworming is a good idea, and that the costs are low relative to other interventions. My main point here is different, that conducting good RCTs is difficult and often expensive, so that problems often arise that need to be dealt with by various econometric or statistical fixes. There is nothing wrong with such fixes in principle—though they often compromise the substance, as in the instrumental variable estimation to correct for failure of assignment—but their application takes us out of the world of ideal RCTs and back into the world of everyday econometrics and statistics, so that RCTs, although frequently useful, are not exempt from the routine statistical and substantive scrutiny that should be routinely applied to any empirical investigation.

Although it is well beyond my scope in this paper, I should note that RCTs in medicine—the gold standard to which development RCTs often compare themselves—are not exempt from practical difficulties, and their primacy is not without challenge. In particular, ethical (human subjects) questions surrounding RCTs in medicine have become sufficiently severe to seriously limit what can be undertaken. At the same time there is much concern that those who sponsor trials and those who analyse them have large financial stakes in the outcome, which sometimes casts doubts on the results. This is not currently a problem in economics, but would surely become one if, as the advocates argue, RCTs became a precondition for funding of projects. Beyond that Concato *et al*. (2000) argue that, in practice, RCTs do not provide useful information beyond what can be learned from well-designed and carefully interpreted observational studies.

5. Where should we go from here?

Cartwright (2007*b*) maintains a useful distinction between 'hunting' causes and 'using' them, and this Section is about the use of randomised controlled trials for policy. This raises the issue of generalisability or external validity—as opposed to internal validity as discussed in the previous section—grounds on which development RCTs are sometimes criticised (see, for example, Rodrik 2008).

There are certainly cases in both medicine and economics where an RCT has had a major effect on the way that people think. In the recent

development literature, my favourite is Chattopadhyay and Duflo's (2004) study of women leaders in India. Some randomly selected *panchayats* were forced to have female leaders, and the paper explores the differences in outcomes between such villages and others with male leaders. There is a theory (of sorts) underlying these experiments; the development community had for a while adopted the view that a key issue in development was the empowerment of women (or perhaps just giving them 'voice') and, if this was done, more children would be educated, more money would be spent on food and on health, and there would be many other socially desirable outcomes. Women are altruistic and men are selfish. Chattopadhyay and Duflo's analysis of the Indian government's experiments shows that this is wrong. There are many similar examples in medicine where knowledge of the mean treatment effect among the treated, even with some allowance for practical problems, is difficult to reconcile with currently held beliefs.

Yet I also believe that RCTs of 'what works', even when done without error or contamination, are unlikely to be helpful for policy unless they tell us something about why it works, something to which they are often neither targeted nor well-suited. Some of the issues are familiar and are widely discussed in the literature. Actual policy is always likely to be different from the experiment, for example, because there are general equilibrium effects that operate on a large scale that are absent in a pilot, or because the outcomes are different when *everyone* is covered by the treatment rather than just a selected group of experimental subjects. Small development projects that help a few villagers or a few villages may not attract the attention of corrupt public officials because it is not worth their while to undermine or exploit them, yet they would do so as soon as any attempt were made to scale up. The scientists who run the experiments are likely to do so more carefully and conscientiously than would the bureaucrats in charge of a full scale operation. So there is no guarantee that the policy tested by the RCT will have the same effects as in the trial, even on the subjects included in the trial.

It is sometimes argued that scepticism about external validity is simply 'a version of David Hume's justly famous demonstration of the lack of a rational basis for induction' (Banerjee 2005). But what is going on here is often a good deal more mundane. Worrall (2007: 995) responds to the same argument with the following:

> One example is the drug benoxaprophen (trade name: Opren), a nonsteroidal inflammatory treatment for arthritis and musculo-skeletal pain. This passed RCTs (explicitly restricted to 18 to 65 year olds) with flying colours. It is how-

> ever a fact that musculo-skeletal pain predominately afflicts the elderly. It turned out, when the (on average older) 'target population' were given Opren, there were a significant number of deaths from hepato-renal failure and the drug was withdrawn.

In the same way, an educational protocol that was successful when randomised across villages in India holds many things constant that would not be constant if the programme were transported to Guatemala or Vietnam. These examples demonstrate a failure to control for relevant factors, not the impossibility of induction.

Pawson and Tilley (1997) argue that it is the combination of mechanism and context that generates outcomes and that without understanding that combination scientific progress is not possible. Nor can we safely go from experiments to policy. In economics, the language would be about theory, building models, and tailoring them to local conditions. Policy requires a causal model; without it, we cannot understand the welfare consequences of a policy, even a policy where causality is established and is proven to work on its own terms. Banerjee (2007*b*) describes an RCT by Duflo, Hanna and Ryan (2008) as 'a new economics being born'. This experiment used cameras to monitor and prevent teacher absenteeism in villages in the Indian state of Rajasthan. Curiously, Pawson and Tilley (1997: 78–82) use the example of cameras (to deter crime in car parks) as one of their running examples. They note that cameras do not, in and of themselves, prevent crime because they do not make it impossible to break into a car. Instead, they depend on triggering a series of behavioural changes. Some of those changes show positive experimental outcomes—crime is down in the car parks with cameras—but are undesirable, for example because crime is shifted to other car parks, or because the cameras change the mix of patrons of the car park. There are also cases where the experiment fails but has beneficial effects. It would not be difficult to construct similar arguments for the cameras in the Indian schools, and welfare conclusions cannot be supported unless we understand the behaviour of teachers, pupils, and their parents. Duflo, Hanna, and Ryan (2008) understand this, and use their experimental results to construct a model of teacher behaviour; other papers that use structural models to interpret experimental results include Todd and Wolpin (2006) and Attanasio, Meghir and Santiago (2005).

Cartwright (2007*a*) draws a contrast between the rigour applied to establish internal validity—to establish the gold standard status of RCTs—and the much looser arguments that are used to defend the transplantation of the experimental results to policy. For example, running

RCTs to find out whether a project works is often defended on the grounds that the experimental project is like the policy that it might support. But the 'like' is typically argued by an appeal to similar circumstances, or a similar environment, arguments that can only be mounted for observable variables. Yet controlling for observables is the key to the matching estimators that are the main competitors for RCTs, and that are typically rejected by their advocates on the grounds that RCTs control not only for things that we observe but things that we cannot. As Cartwright notes, the validity of evidence-based policy depends on the *weakest* link in the chain of argument and evidence, so that by the time we seek to use the experimental results, the advantage of RCTs over matching or other econometric methods has evaporated. In the end, there is no substitute for careful evaluation of the chain of evidence and reasoning by people who have the experience and expertise in the field. The demand that experiments be theory-driven is, of course, no guarantee of success, though the lack of it is close to a guarantee of failure.

It is certainly not always obvious how to combine theory with experiments. Indeed, much of the interest in RCTs—and in instrumental variables and other econometric techniques that mimic random allocation—comes from a deep scepticism about much economic theory, and impatience with its ability to deliver structures that seem at all helpful in interpreting reality. The wholesale abandonment in American graduate schools of price theory in favour of infinite horizon intertemporal optimisation and game theory has not been a favourable development for young empiricists. Empiricists and theorists seem further apart now than at any period in the last quarter century. Yet reintegration is hardly an option because without it there is no chance of long term scientific progress. One promising area is the recent work in behavioural economics, and the closer integration of economics and psychology, whose own experimental tradition is clearly focused on behavioural regularities. The experiments reviewed in Levitt and List (2009), often involving both economists and psychologists, cover such issues as loss aversion, procrastination, hyperbolic discounting and the availability heuristic, all of which are examples of behavioural mechanisms that promise applicability beyond the specific experiments. There also appears to be a good deal of convergence between this line of work, inspired by earlier experimental traditions in economic theory and in psychology, and the most recent work in development. Instead of using experiments to evaluate projects, looking for which projects work, this development work designs its experiments to test predictions of theories that are generalisable to other situ-

ations. Without any attempt to be comprehensive, some examples are Karlan and Zinman (2008), who are concerned with the price-elasticity of the demand for credit, Bertrand *et al*. (forthcoming), who take predictions about the importance of context from the psychology laboratory to the study of advertising for small loans in South Africa, Duflo, Kremer and Robinson (2009), who construct and test a behavioural model of procrastination for the use of fertilisers by small farmers in Kenya, and Giné and Karlan (2008), who use an experiment in the Philippines to test the efficacy of a smoking-cessation product designed around behavioural theory. In all of this work, the project, when it exists at all, is the embodiment of the theory that is being tested and refined, not the object of evaluation in its own right, and the field experiments are a bridge between the laboratory and the analysis of 'natural' data (List 2006). The collection of purpose-designed data and the use of randomisation often make it easier to design the sort of acid test that can be more difficult to construct without them. If we are lucky, this work will provide the sort of behavioural realism that has been lacking in much of economic theory while, at the same time, identifying and allowing us to retain the substantial parts of existing economic theory that remain genuinely useful.

In this context, it is worth looking back to the previous phase of experimentation in economics that started with the New Jersey income tax experiments. A rationale for these experiments is laid out in Orcutt and Orcutt (1968), in which the vision is a formal model of labour supply with the experiments used to estimate its parameters. By the early 1990s, however, experimentation had moved on to a 'what works' basis and Manski and Garfinkel (1992), surveying the experience, write 'there is, at present, no basis for the popular belief that extrapolation from social experiments is less problematic than extrapolation from observational data. As we see it, the recent embrace of reduced-form social experimentation to the exclusion of structural evaluation based on observational data is not warranted.' Their statement still holds good, and it would be worth our while trying to return to something like Orcutt and Orcutt's vision.

Finally, I want to return to the issue of 'heterogeneity', a running theme in this lecture. Heterogeneity of responses first appeared in Section 2 as a technical problem for instrumental variable estimation, dealt with in the literature by local average treatment estimators. Randomised controlled trials provide a method for estimating quantities of interest in the presence of heterogeneity, and can therefore be seen as another technical solution for the 'heterogeneity problem'. They allow estimation of mean

responses under extraordinarily weak conditions. But as soon as we deviate from ideal conditions, and try to correct the randomisation for inevitable practical difficulties, heterogeneity again rears its head, biasing estimates, and making it difficult to interpret what we get. In the end, the technical fixes fail and compromise our attempts to learn from the data. What this should tell us is that the heterogeneity is *not* a technical problem, but a symptom of something deeper, which is the failure to specify causal models of the processes we are examining. This is the methodological message of this lecture, that technique is never a substitute for the business of doing economics.

Note. I am grateful to Abhijit Banerjee, Tim Besley, Anne Case, Hank Farber, Bill Easterly, Bo Honoré, Michael Kremer, David Lee, Chris Paxson, Sam Schulhofer-Wohl, Burt Singer, Jesse Rothstein, and John Worrall for helpful discussions in the preparation of this paper. For comments on a draft, I would like to thank Tim Besley, Richard Blundell, David Card, Winston Lin, John List, Costas Meghir, David McKenzie, Burt Singer, Alessandro Tarozzi, Gerard van den Berg, Eric Verhoogen and especially Esther Duflo. I would also like to acknowledge a particular intellectual debt to Nancy Cartwright, who has discussed these issues patiently with me for several years, whose own work on causality has greatly influenced me, and who pointed me towards other important work; to Jim Heckman, who has long thought deeply about the issues in this lecture, and many of whose views are recounted here; and to David Freedman, whose recent death has deprived the world of one of its greatest statisticians, and who consistently and effectively fought against the (mis)use of technique as a substitute for substance and thought. None of which removes the need for the usual disclaimer, that the views expressed here are entirely my own. I acknowledge financial support from NIA through grant P01 AG05842–14 to the National Bureau of Economic Research.

References

Acemoglu, D., Johnson, S. and Robinson, J. A. (2001), 'The colonial origins of comparative development: an empirical investigation', *American Economic Review*, 91 (5): 1369–1401.

Altman, D. G. (1998), 'Within trial variation—a false trail', *Journal of Clinical Epidemiology*, 51 (4): 301–3.

Angrist, J. D. (1990), 'Lifetime earnings and the Vietnam era draft lottery', *American Economic Review*, 80 (3): 313–36.

Angrist, J. D. and Imbens, G. (1994), 'Identification and estimation of local average treatment effects', *Econometrica*, 62 (2): 467–75.

Angrist, J. D. and Krueger, A. (1991), 'Does compulsory school attendance affect schooling and earnings?', *Quarterly Journal of Economics*, 106 (4): 979–1014.

Angrist, J. D. and Lavy, V. (1999), 'Using Maimonides' rule to estimate the effect of class size on scholastic achievement', *Quarterly Journal of Economics*, 114 (2): 533–75.

Attanasio, O., Meghir, C. and Santiago, A. (2005), 'Education choices in Mexico: using a structural model and a randomized experiment to evaluate Progresa' (London, Institute for Fiscal Studies, processed).

Banerjee, A. V. (2005), '"New development economics" and the challenge to theory', *Economic and Political Weekly*, 1 October, 4340–4.

Banerjee, A. V. (2007*a*), *Making Aid Work* (Cambridge, MA).

Banerjee, A. V. (2007*b*), 'Inside the machine: toward a new development economics', *Boston Review*, 4 September.

Banerjee, A. V. and He, R. (2008), 'Making aid work', in William R. Easterly (ed.), *Reinventing Foreign Aid* (Cambridge, MA), pp. 47–92.

Barro, R. J. (1998), *Determinants of Economic Growth: a Cross-Country Empirical Study* (Cambridge, MA).

Barro, R. J. and Sala-i-Martin, X. (1995), *Economic Growth* (New York).

Bauer, T. P. (1971), *Dissent on Development: Studies and Debates in Development Economics* (London).

Bauer, T. P. (1981), *Equality, the Third World, and Economic Delusion* (Cambridge, MA).

Bertrand, M., Karlan, D. S., Mullainathan, S., Shafir, E. and Zinman, J. (forthcoming), 'What's advertising content worth? Evidence from a consumer credit marketing field experiment' *Quarterly Journal of Economics*.

Boone, P. (1996), 'Politics and the effectiveness of foreign aid', *European Economic Review*, 40: 289–329.

Burnside, C. and Dollar, D. (2000), 'Aid, policies, and growth', *American Economic Review*, 90 (4): 847–67.

Cartwright, N. (2007*a*), 'Are RCTs the gold standard?', *Biosocieties*, 2: 11–20.

Cartwright, N. (2007*b*), *Hunting Causes and Using Them: Approaches in Philosophy and Economics* (Cambridge).

Chattopadhyay, R. and Duflo, E. (2004), 'Women as policy makers: evidence from a randomized controlled experiment in India', *Econometrica*, 72 (5): 1409–43.

Clemens, M., Radelet, S. and Bhavnani, R. (2004), 'Counting chickens when they hatch: the short-term effect of aid on growth', Center for Global Development, Working Paper 44, November. http://www.cgdev.org/content/publications/detail/2744 (accessed 16 December 2008).

Concato, J., Shah, N. and Horwitz, R. I. (2000), 'Randomized, controlled trials, observational studies, and the hierarchy of research designs', *New England Journal of Medicine*, 342 (25): 1887–92.

Cox, D. R., 1958, *Planning of Experiments*, New York, Wiley.

Dalgaard, C.-J. and Hansen, H. (2001), 'On aid, growth, and good policies', *Journal of Development Studies*, 37 (6): 17–41.

Dalgaard, C.-J., Hansen, H. and Tarp, F. (2004), 'On the empirics of foreign aid and growth', *Economic Journal*, 114: F191–216.

Davey Smith, G. and Egger, M. (1998), 'Incommunicable knowledge? Interpreting and applying the results of clinical trials and meta-analysis', *Journal of Clinical Epidemiology*, 51 (4): 289–95.

De Mel, S., McKenzie, D. and Woodruff, C. (2008), 'Returns to capital in micro-enterprises: evidence from a field experiment', *Quarterly Journal of Economics*, 123 (4): 1329–72.

Duflo, E., 2004, 'Scaling up and evaluation', *Annual World Bank Conference on Development Economics 2004* (Washington, DC: The World Bank).

Duflo, E. and Pande, R. (2007), 'Dams', *Quarterly Journal of Economics*, 122 (2): 601–46.

Duflo, E., Glennerster, R. and Kremer, M. (2008), 'Using randomization in development economics research: a toolkit', Chapter 61 in T. Paul Schultz and John Strauss (eds.), *Handbook of Development Economics*, Vol. 4 (Amsterdam), pp. 3895–962.

Duflo, E., Hanna, R. and Ryan, S. (2008), 'Monitoring works: Getting teachers to come to school', CEPR Working Paper No. 6682, February.

Duflo, E., Kremer, M. and Robinson, J. (2009), 'Nudging farmers to use fertilizer: evidence from Kenya' (MIT, processed).

Easterly, W. R. (2006), *The White Man's Burden: Why the West's Efforts to Aid the Rest Have Done So Much Ill and So Little Good* (Oxford).

Easterly, W. R. (ed.) (2008), *Reinventing Foreign Aid* (Cambridge, MA).

Easterly, W. R., Levine, R. and Roodman, D. (2003), 'Aid, policies, and growth: Comment', *American Economic Review*, 94 (3): 774–80.

Feins, J. D. and McInnis, D. (2001), 'The interim impact evaluation for the moving to opportunity demonstration' (Abt Associates), http://www.nber.org/~kling/mto/MTO_OMB.pdf (accessed 15 December 2008).

Fisher, R. A. (1935), *The Design of Experiments* (8th edn., 1960) (New York).

Freedman, D. A. (2004), *Statistical Models: Theory and Practice* (New York).

Freedman, D. A. (2006), 'Statistical models for causation: what inferential leverage do they provide?', *Evaluation Review*, 30 (6): 691–713.

Freedman, D. A. (2008), 'On regression adjustments to experimental data', *Advances in Applied Mathematics*, 40: 180–93.

Giné, X. and Karlan, D. S. (2008), 'Put your money where your butt is: a commitment contract for smoking cessation' (Yale University, processed).

Glewwe, P., Kremer, M., Moulin, S. and Zitzewitz, E. (2004), 'Retrospective vs. prospective analyses of school inputs: the case of flip charts in Kenya', *Journal of Development Economics*, 74: 251–68.

Guillaumont, P. and Chauvet, L. (2001), 'Aid and performance: a reassessment', *Journal of Development Studies*, 37 (6): 66–92.

Hansen, H. and Tarp, F. (2000), 'Aid effectiveness disputed', *Journal of International Development*, 12: 375–98.

Hansen, H. and Tarp, F. (2001), 'Aid and growth regressions', *Journal of Development Economics*, 64: 547–70.

Heckman, J. J. (1992), 'Randomization and social program evaluation', in Charles Manski and Irwin Garfinkel (eds.), *Evaluating Welfare and Training Programs* (Cambridge, MA). pp. 201–30. (Available as NBER Technical Working Paper No. 107.)

Heckman, J. J. (1997), 'Instrumental variables: a study of implicit behavioral assumptions used in making program evaluations', *The Journal of Human Resources*, 32 (3): 441–62.

Heckman, J. J. (2000), 'Causal parameters and policy analysis in economics: a twentieth century retrospective', *Quarterly Journal of Economics*, 115: 45–97.

Heckman, J. J. and Smith, J. A. (1995), 'Assessing the case for social experiments', *Journal of Economic Perspectives*, 9 (2): 85–115.

Heckman, J. J. and Vytlacil, E. J. (1999), 'Local instrumental variables and latent variable models for identifying and bounding treatment effects', *Proceedings of the National Academy of Sciences*, 96(8), 4730–4.

Heckman, J. J. and Vytlacil, E. J. (2007), 'Econometric evaluation of social programs, Part 2: Using the marginal treatment effect to organize alternative econometric estimators to evaluate social programs, and to forecast their effects in new environments', Chapter 71, in James J. Heckman and Edward E. Leamer (eds.), *Handbook of Econometrics*, *Volume 6B* (Amsterdam), pp. 4875–5143.

Heckman, J. J., Urzua, S. and Vytlacil, E. J. (2006), 'Understanding instrumental variables in models with essential heterogeneity', *Review of Economics and Statistics*, 88 (3): 389–432.

Horwitz, R. I., Singer, B. M., Makuch, R. W. and Viscoli, C. M. (1996), 'Can treatment that is helpful on average be harmful to some patients? A study of the conflicting information needs of clinical inquiry and drug regulators', *Journal of Clinical Epidemiology*, 49: 395–400.

Horwitz, R. I., Singer, B. M., Makuch, R. W. and Viscoli, C. M. (1997), 'On reaching the tunnel at the end of the light', *Journal of Clinical Epidemiology*, 50 (7): 753–55.

Hoxby, C. M. (2000), 'Does competition among public schools benefit students and taxpayers?', *American Economic Review*, 90 (5): 1209–38.

International Initiative for Impact Evaluation (3IE) (2008), http://www.3ieimpact.org/.

Kanbur, R. M. (2001), 'Economic policy, distribution, and poverty: the nature of the disagreements', *World Development*, 29 (6): 1083–94.

Karlan, D. S. and Zinman, J. (2008), 'Credit elasticities in less developed economies: implications for micro-finance', *American Economic Review*, 98 (3): 1040–68.

Leamer, E. E. (1985), 'Vector autoregressions for causal inference?', *Carnegie-Rochester Conference Series on Public Policy*, 22: 255–304.

Lensink, R. and White, H. (2001), 'Are there negative returns to aid?', *Journal of Development Studies*, 37 (6): 42–65.

Leonhardt, D. (2008), 'Making economics relevant again', *New York Times*, 20 February.

Levitt, S. D. and List, J. A. (2009), 'Field experiments in economics: the past, the present, and the future', *European Economic Review*, 53 (1): 1–18.

List, J. A. (2006), 'Field experiments: a bridge between lab and naturally occurring data', *Advances in Economic Analysis and Policy*, 6 (2): 1–45.

Mankiw, N. G., Romer, D. and Weil, D. N. (1992), 'A contribution to the empirics of economic growth', *Quarterly Journal of Economics*, 107 (2): 407–37.

Manski, C. and Garfinkel, I. (1992), 'Introduction', in Charles Manski and Irwin Garfinkel (eds.), *Evaluating Welfare and Training Programs* (Cambridge, MA), pp.1–22.

McCleary, R. M. and Barro, R. J. (2006), 'Religion and economy', *Journal of Economic Perspectives*, 20 (2): 49–72.

Miguel, E. and Kremer, M. (2004), 'Worms', *Econometrica*, 72 (1): 159–217.

Miguel, E., Satyanath, S. and Sergenti, E. (2004), 'Economic shocks and civil conflict: an instrumental variables approach', *Journal of Political Economy*, 112 (4): 725–53.

Orcutt, Guy H. and Orcutt, Alice G. (1968), 'Incentive and disincentive experimentation for income maintenance purposes', *American Economic Review*, 58 (4): 754–72.

Pogge, T. (2005), 'World poverty and human rights', *Ethics and International Affairs*, 19 (1): 1–7.

Poverty Action Lab (2007), 'Clinton honors global deworming effort', http://www.povertyactionlab.org/deworm/ (accessed 15 December 2008).

Pawson, R. and Tilley, N. (1997), *Realistic Evaluation* (London).

Rajan, R. and Subramanian, A. (2005), 'Aid and growth: what does the data really show?', NBER Working Paper No. 11513, June. *Review of Economics and Statistics* (forthcoming).

Rodrik, D., 2008, 'The new development economics: we shall experiment, but how shall we learn?' (July), http://ksghome.harvard.edu/~drodrik/The%20New%20Development%20Economics.pdf (accessed 15 December 2008).

Roodman, D. (2008), 'The anarchy of numbers: aid, development, and cross-country empirics', *World Bank Economic Review*, 21 (2): 255–77.

Sachs, J. (2005), *The End of Poverty: Economic Possibilities for Our Time* (New York).

Sachs, J. (2008), *Common Wealth: Economics for a Crowded Planet* (New York).

Senn, S. and Harrell, F. (1997), 'On wisdom after the event', *Journal of Clinical Epidemiology*, 50 (7): 749–51.

Singer, P. (2004), *One World: the ethics of globalization*, 2nd edn. (New Haven).

The Lancet (2004), 'The World Bank is finally embracing science', Editorial, 364, 28 August, 731–2

Todd, P. E. and Wolpin, K. I. (2006), 'Assessing the impact of a school subsidy program in Mexico: using a social experiment to validate a dynamic behavioral model of child schooling and fertility', *American Economic Review*, 96 (5): 1384–1417.

Urquiola, M. and Verhoogen, E. (2009), 'Class-size caps, sorting, and the regression-discontinuity design', *American Economic Review*, 99 (1): 179–215.

van den Berg, G. (2008), 'An economic analysis of exclusion restrictions for instrumental variable estimation' (VU University Amsterdam, processed).

World Bank (2008), The Spanish Impact Evaluation Fund, http://web.worldbank.org/WBSITE/EXTERNAL/TOPICS/EXTPOVERTY/EXTISPMA/0,,contentMDK:21419502~menuPK:384336~pagePK:148956~piPK:216618~theSitePK:384329,00.html (accessed 15 December 2008).

Worrall, J. (2007), 'Evidence in medicine and evidence-based medicine', *Philosophy Compass*, 2 (6): 981–1022.