Biology and Philosophy (2006) 21:253–276
DOI 10.1007/s10539-005-8304-3

Causal efficacy and the analysis of variance

ROBERT NORTHCOTT
University of Missouri-St. Louis, 16 Fawley Road, NW6 1SH, London, United Kingdom (e-mail: northcottr@usml.edu)

Received 22 July 2004; accepted in revised form 2 June 2005

Key words: Analysis of variance, Casual efficacy, Counterfactuals, Environment, Genes, Nature, Nurture

Abstract. The causal impacts of genes and environment on any one biological trait are inextricably entangled, and consequently it is widely accepted that it makes no sense in singleton cases to privilege either factor for particular credit. On the other hand, at a population level it may well be the case that one of the factors is responsible for more variation than the other. Standard methodological practice in biology uses the statistical technique of *analysis of variance* to measure this latter kind of causal efficacy. In this paper, I argue that:
1) analysis of variance is in fact badly suited to this role; and
2) a superior alternative definition is available that readily reconciles *both* the entangled-singleton *and* the population-variation senses of causal efficacy.

Causal efficacy

Is a plant's height due more to genes or environment? Is mutation, migration or genetic drift a more important force in evolution? Was I delayed more by the roadworks or by the multiple traffic lights? The idea of causal efficacy, or of some causes being more or less important than others, is ubiquitous in biology and indeed much else of science and everyday life. It might seem that the notion is pretty straightforward: can we not just define the efficacy of a cause by the quantity of effect it leads to? But it will turn out that the task is less simple than it first appears, and moreover that the leading definition of causal efficacy in use in biology – namely the statistical technique of analysis of variance – is, I shall argue, an example of a method fundamentally ill suited to its purpose.

Notice from the start that the issue here is *not* the venerable philosophical project of defining causation itself. Neither is the issue related to that of how best to infer causes from statistical data. Instead, I shall assume always that – as is common in practical problems – all parties already agree on what causes are present. The issue at hand is, rather, those causes' relative importance.

The philosophical literature specifically on causal efficacy is relatively sparse.[1] This may be in part because in some well-known contexts there does not seem to be much of interest to say. Take a Newtonian particle, for example, and the question of whether its acceleration is due more to gravity or to electricity. We understand easily enough that gravity's causal efficacy is given by the extra acceleration its presence imparts to the particle, and likewise for electricity. There seems to be no problem or ambiguity with this understanding, and moreover it is straightforward then to compare the two influences' relative importance.

But turn now from Newtonian physics to the nature–nurture issue in biology. Unfortunately, it quickly becomes clear that weighing up the relative importance of genes and environment is a more complex matter than it was for gravity and electricity. Gravity may accelerate some particle 10% more than does electricity but can we make sense of claiming, say, that genes produced some person's high exam scores 10% more than did environment? The impacts of the genetic and environmental factors are of course hopelessly interwoven from the start and there is none of the easy separability of the physics case. Without the right environmental input from the womb onwards, even Einsteinian genes will not produce high exam scores. And likewise without genes to work on, even the best nutrients and most scholarly philosophy department will be unable to produce any high exam scores either. Given this interdependence, or more precisely that each input is insufficient to produce any final effect without the other, at first sight it seems difficult to disentangle any causal efficacies at all. And this indeed is precisely the professional consensus among biologists – to speak of genes or environment having particular efficacies is, in this context, meaningless. However, all is not completely lost since in response to these conceptual difficulties biologists have developed instead a slightly different understanding of causal efficacy.

The analysis of variance

Suppose a farmer wishes to maximise the heights, and hence yields, of some corn plants. She has available a range of three different plant strains (say Cheap, Normal and Expensive) and also a range of three different fertiliser treatments (Green, FastGro and SuperGro). She can thus generate a total of nine different gene-environment combinations, corresponding to each possible combination of strain and fertiliser. Suppose finally that we plot a table of (average) plant heights for each of these combinations, yielding the results in Table 1.

Some fertilisers achieve better results than do others; likewise so do some plant strains. But which of the two factors, genes or environment, is the more important? Intuitively, it is easy to grasp the following argument: given any

[1] Rare discussions include those in Good (1961), Miller (1987), Sober (1988) – from which my Newtonian particle and corn plant examples will be adapted – Sober et al. (1992), Pearl (2000) and Spirtes et al. (2000).

Table 1. Fictional data for plant heights and gene-environment combinations.

	SuperGro	FastGro	Green	MA
Expensive	145	115	85	115
Normal	140	110	80	110
Cheap	135	105	75	105
MA	140	110	80	grand mean: 110

'MA' stands for marginal average. For simplicity assume, both here and later in Table 3, that each cell has equal weighting.

particular fertiliser, choice of genes makes only a small difference, outcomes ranging only over 10 cm. By contrast, given any particular plant strain, choice of environment makes a huge difference, yielding a 60-cm range of plant heights. That is, choice of environment matters much more than choice of genes. In this population, changing plant strains would be a bad way to increase plant heights – much better to concentrate instead on good fertilisers since the latter factor makes so much more difference.

This kind of argument is the essence of the alternative understanding of causal efficacy, whereby comparison of genes and environment need no longer be declared inevitably hopeless or meaningless. On the contrary, as we just saw, there is a clear sense in which one of them can be deemed a stronger cause than the other. The actual formal definition borrows from the statistical technique of analysis of variance ('ANOVA').[2] In this example, it would proceed first by calculating the sum of squares ('SS') across the different plant strains' marginal averages, i.e. the sum of squares of the figures in the last column.[3] Thus:

$$\text{Genes SS} = 3[(115-110)^2 + 0^2 + (105-110)^2] = 3(25+0+25) = 150$$

(The multiplying coefficient 3 is a function of the table being 3×3, but we need not go into this detail of ANOVA here.) Similarly the formal definition would next calculate the sum of squares across the different fertilisers' marginal averages, i.e. across the figures of the bottom row:

$$\begin{aligned}\text{Environment SS} &= 3[(140-110)^2 + (110-110)^2 + (80-110)^2] \\ &= 3(900+0+900) = 5400\end{aligned}$$

Finally, it would then compare the two SS figures to see which factor accounted for a greater proportion of the total variation in the population. In this case, $5400 > 150$, so more of the variation would be 'explained' by choice of fertiliser than choice of plant strain, and thus environment be awarded a higher causal efficacy than genes. For our purposes, the further mathematical

[2] I provide only a simplified account – see Sokal et al. (1995), or any other standard textbook, for more details. Note that our concern is not with ANOVA inference, hence the omission here of all reference to error terms.

[3] In accordance with standard usage, by 'sum of squares' is meant taking each figure's deviation from the mean, squaring that deviation, and then adding together all these squared deviations.

details of the technique do not matter; what is important is to appreciate the attractiveness of the intuition underpinning it.

Thus biologists, once *given* a range of values for each input, have a way of defining how varying one of them can be said to have more impact on the effect term than does varying the other. A couple of features of the technique are noteworthy here. First, the values ANOVA yields for causal efficacies are critically dependent on choice of population. Perhaps for a different range of plant strains (Ancient, Normal, and GM) and fertilisers (Slightly Below Average, Average, and Slightly Above Average) the results would have been very different, indeed reversed: now we might have found that choice of genes was responsible for more of the variation in plant height than was choice of environment. Lewontin (1974) demonstrates how misleading it can be to extrapolate ANOVA results uncritically. In particular, they cannot reliably predict the outcomes of interventions that involve levels of input outside the range of the generating population. The second noteworthy feature is that, ANOVA does not, nor is claimed to, explain the occurrence of *singleton* effects. For instance it does not tell us which of genes and environment explained more of some individual Expensive-SuperGro corn plant's height of 145 cm. (One reason is that ANOVA takes every cell's score into account, yet the height of a plant in any one cell of the table seems clearly to be causally independent of the height in any other cell.) Rather, it only assigns responsibility for the pattern of outcomes across a particular *population*, i.e. assigns causal efficacies only at the group rather than singleton level.[4]

It might be worth emphasising more why ANOVA must indeed be seen as a proxy measure of causal efficacy. Its own procedure makes explicit reference to the distinction between independent and dependent variables, the latter being seen as a function of the former rather than the other way round – an asymmetry already obviously redolent, if not directly derivative, of that between cause and effect. It also yields a measure of strength of association between two variables. When one of these variables is taken to be the cause of the other, such a measure is normally (indeed can hardly not be) interpreted as one of causal *efficacy* – how *much*, or how *strong*, an impact does the cause variable have on the effect variable? Thus Remington and Schork (1970: 229, my italics) comment, for instance: "the main objective of the analysis of variance is to assess the *influence* of each factor … upon some response variable."

To make this claim about the use of ANOVA seem plausible in the space available here, consider the actual subjects of ANOVA study (Table 2) [5]

In each case the choice of variables clearly suggests particular cause and effect relations and so any results linking them quantitatively ought indeed to be interpreted as (claimed) causal efficacies. I shall argue, however, that using ANOVA in this way, although thus ubiquitous and apparently reasonable, is

[4] Later I consider, but reject, a way of extending ANOVA to the singleton case.

[5] These are studies chosen by standard textbooks precisely to illustrate appropriate usage. I take them therefore also to be cases of reputable work and not unrepresentatively sloppy. The textbooks themselves (Sokal and Rohlf 1981, 1995; Bliss 1967) contain the individual references.

Table 2. Topics of actual 2-factor ANOVAs in biology.

1st independent variable (first presumed cause)	2nd independent variable (second presumed cause)	Dependent variable (presumed effect)
Type of sugar treatment	pH level	Growth of peas
Density per container	Three different strains	Housefly's developmental period
Level of thyroxin injected	Sex of chick	Weight at 7 weeks
Month of year	14 different varieties	Degree of aphid infestation of potato tubers
Depth	Day of measurement	A lake's water temperature
Different locations	Surface soil or subsoil	Soil pH
Breed of cow	Age of cow	Butterfat percentage
Nitrogen concentration	Variety of wood chips added	Yield of beet roots
Length of frost-free season	Amount of light	Height of hemlock buds
quantity of nitrogen fertiliser	quantity of phosphoric acid fertiliser	yield of corn
sodium concentration	potassium concentration	weight of tobacco leaves

nevertheless mistaken. To demonstrate why, it will be necessary first to return at greater length to fundamentals.

A unified account of causal efficacy

To summarise so far: in physics, it seems there are no problems or ambiguities. In biology by contrast, at least in the case of nature-nurture, one understanding of causal efficacy seems to render the notion meaningless. However, a more successful second understanding of it can still be achieved, albeit only indirectly via the technique of analysis of variance across a population. Sober (1988: 304) concludes from this that: "there is no such thing as the way science apportions causal responsibility; rather, we must see how different sciences understand this problem differently." But I want to reject that conclusion. My opinion, on the contrary, is that a unified account across sciences of causal efficacy *is* possible, and that the apparently divergent cases of physics and biology in fact turn out merely to be different manifestations of a common underlying principle.

A definition

Let the *efficacy* of a cause C with respect to an effect E be:

$$E(C \,\&\, W_1) - E(C_0 \,\&\, W_0)$$

where W_1 and W_0 are background conditions (on which more shortly), and C_0 is a baseline counterfactual level of C. In the simplest case, C_0 will just be the absence[6]

[6] Strictly speaking, of course, both here and later 'absence' should be thought of merely as convenient shorthand for some alternative event, for instance for keeping a fertiliser in its container rather than spraying it on the crop.

of C. More generally, often it will be useful to set C_0 to be at what Humphreys (1989: 38) calls the *neutral level* of C. This he defines, in the case of a variable, as "the level of the variable at which the property corresponding to that variable is completely absent." For instance, a neutral level of fertiliser in our example might be one that, if added, would make no difference to the corn plant's height. A key point is that this neutral level is likely to depend on the exact effect variable and the exact circumstances of interest.

The intuition behind the definition is straightforward: we are interested in the quantity of effect for which C is responsible, and this is just the level of effect with C compared to what that level would have been without it. For example, the efficacy of a kick of a ball is yielded by that ball's acceleration compared to its acceleration if it had *not* been kicked. Any value for the efficacy of C will be relativised both to the levels of other causes of E (reflected in the background conditions W) and also to the choice of C_0. In essence, the first of these relativisations captures an efficacy's dependence on choice of W. For example, striking a match will cause a light if background conditions include sufficient oxygen in the atmosphere but not otherwise. The second relativisation captures the way in which the efficacy ascribed to a cause also depends on the choice of contrast class to which we compare it. For instance, the gain from kicking a ball rather than throwing it would be different from the gain of kicking it rather than leaving it untouched.

What is essential is that background conditions W are kept appropriately *constant* across the two terms. This captures the logic of controlled experiment – to assess the impact of introducing C in place of C_0, we want in an experiment to keep constant (as best we can) everything else causally relevant. Note though that, strictly speaking, the W in the left-hand term is different to that in the right-hand term, because the switch between C and C_0 will in general alter additional things in the world besides its impact on our effect of interest. It turns out that the definition's exact formulation also depends on a number of other technicalities, but for our purposes we may gloss over those without loss.[7]

Generally, the definition is clearly redolent of the counterfactual analysis of causation itself, first put forward in its modern form by Lewis (1973). Note that our definition here appeals to a specific counterfactual contrast class C_0, and also that it is able to offer quantitative results. In both these respects it owes much also to the literature on probabilistic causality. A virtue of the formulation is its clear applicability to scientific practice, in particular the connection to controlled experiment mentioned above. (This virtue is also emphasised by manipulationist versions of the counterfactual view: Woodward 2003.) For now, I leave a fuller discussion of how the relevant counterfactuals might actually be evaluated until Section 7. Before then I shall simply assume in my examples that such evaluations are unproblematic, although see also footnote 11.

It might seem that the definition is so general as to be everywhere rather obvious and therefore not particularly helpful, but this turns out not to be so.

[7] Most notably, extension of the definition to probabilistic rather than deterministic causation. Without loss of generality, I shall also assume in my examples that all cause and effect terms are variables.

Indeed later we shall find an example of actual scientific practice – namely biologists' use of the analysis of variance – that falls foul of it.

Absolute vs. relative causal efficacy

For particular causes C and D, often we are concerned with the following kind of question: 'how much difference did it make having C instead of D?', or 'how much difference did it make having C at level C_1 instead of level C_2?' Start with the latter. Our definition in fact already covers this case, save with respect to C_0 instead of C_2. So we can immediately write that: the causal efficacy of C_1 *relative to* C_2 is:

$$E(C_1 \,\&\, W_1) - E(C_2 \,\&\, W_2)$$

(Let W_2 here be the background conditions that hold given C_2.)

Consider what the causal efficacies of C_1 and C_2 would be instead relative to the neutral level C_0. From before:

(1) Causal efficacy of $C_1 = E(C_1 \,\&\, W_1) - E(C_0 \,\&\, W_0)$
(2) Causal efficacy of $C_2 = E(C_2 \,\&\, W_2) - E(C_0 \,\&\, W_0)$

We see that the *relative* efficacy of C_1 (relative, that is, to C_2) can also be obtained simply by subtracting the two 'absolute' efficacies:

$$\begin{aligned}\text{Causal efficacy of } C_1 \text{ relative to } C_2 &= E(C_1 \,\&\, W_1) - E(C_2 \,\&\, W_2)\\ &= [E(C_1 \,\&\, W_1) - E(C_0 \,\&\, W_0)]\\ &\quad - [E(C_2 \,\&\, W_2) - E(C_0 \,\&\, W_0)]\\ &= (\text{efficacy of } C_1) - (\text{efficacy of } C_2)\end{aligned}$$

(Analogously, the *speed* of one object relative to another is just given by the difference between those two objects' absolute speeds.)

Now turn to the case with two distinct causes C and D: 'how much difference did it make having C instead of D?' Letting C_0 and D_0 be the neutral levels of C and D, respectively, and W_1 and W_2 the appropriately defined background conditions, this can be expressed:

$$E(C \,\&\, W_1 \,\&\, D_0) - E(D \,\&\, W_2 \,\&\, C_0).^{8}$$

Working through our formulas, letting W_0 be the background conditions given C_0 & D_0:

(1) Causal efficacy of C given the absence of D
$= E(C \,\&\, D_0 \,\&\, W_1) - E(C_0 \,\&\, D_0 \,\&\, W_0)$
(2) Causal efficacy of D given the absence of C
$= E(C \,\&\, D_0 \,\&\, W_2) - E(C_0 \,\&\, D_0 \,\&\, W_0)$

[8] Note that we would naturally be interested only in the levels of effect given each cause, each time calculated given the *absence* of the other cause. Else – had we instead calculated each time given the other cause's presence – the efficacy of C relative to D would just have been (for appropriate W_3) trivially zero: E(C & W_3 & D)–E(D & W_3 & C).

We can now again derive a result reducing two causes' relative efficacy just to the difference of their absolute ones (each calculated given the other cause's absence). Thus:

$$\begin{aligned}&\text{Causal efficacy of C relative to D}\\&= E(C \,\&\, W_1 \,\&\, D_0) - E(D \,\&\, W_2 \,\&\, C_0) = [E(C \,\&\, W_1 \,\&\, D_0) - E(C_0 \,\&\, W_0 \,\&\, D_0)]\\&\quad - [E(D \,\&\, W_2 \,\&\, C_0) - E(C_0 \,\&\, W_0 \,\&\, D_0)] = (\text{efficacy of C}) - (\text{efficacy of D})\end{aligned}$$

It follows that whenever we are speaking of a relative causal efficacy we could without loss equally well speak instead of the difference between two absolute efficacies. In other words, strictly speaking the concept of relative causal efficacy is *redundant*.[9]

An example: Holmes and Moriarty

Holmes and Watson finally confront Moriarty, Holmes draws his revolver, and shoots him dead. How important a cause of Moriarty's death was Holmes's shot? 'Very important' would seem to be the obvious answer. But suppose Watson too had a revolver, and that if Holmes had not already done so then Watson would have shot Moriarty himself. It can be argued now that Holmes's shot actually made no difference since whether or not he personally fires, either way Moriarty ends up dead. So it seems there are actually two distinct understandings of causal efficacy: according to one of them Holmes's shot was important, while according to the other it was not.[10]

Here the univocality of causal efficacy now appears doubtful again, but let us analyse the example using the formulas we have just worked out. Label E to be Moriarty's death, C to be Holmes's shot, C_0 to be (some event instantiating) the absence of Holmes's shot, D to be Watson's shot, and D_0 to be (some event instantiating) the absence of Watson's shot. For convenience, let $E = 1$ denote Moriarty's death, and $E = 0$ his survival. Assume that in the actual world Holmes fired but Watson did not.

Begin with the first intuition – that Holmes's shot was indeed an important cause of Moriarty's death. Our formula for the efficacy of Holmes's shot C (in the absence of Watson's shot) is:

$$E(C \,\&\, W_1 \,\&\, D_0) - E(C_0 \,\&\, W_0 \,\&\, D_0)$$

$E(C \,\&\, W_1 \,\&\, D_0) = 1$, since Moriarty dies if Holmes shoots. The right-hand term $E(C_0 \,\&\, W_0 \,\&\, D_0)$ is the counterfactual of if Holmes had not shot (and

[9] It also turns out that an absolute efficacy can always be re-expressed in terms of relative ones, so perhaps one might equally argue that it is absolute efficacy that is redundant. I do not address that issue here. The important point for our purposes is that, whichever formulation is chosen, there is only one independent sense of causal efficacy in play.

[10] This version of the story, and the suggested distinction between two senses of causal efficacy, is taken from Sober (1988).

neither had Watson) – in which case Moriarty would survive, of course, so $E(C_0 \& W_0 \& D_0) = 0$. Therefore, according to our formula, the efficacy of Holmes's shot = 1 – 0 = 1. In other words, Holmes's shot was indeed a maximally important cause.

Turn to the second intuition – that if Watson was going to shoot anyway then Holmes's shot, although indeed the actual cause of Moriarty's death, nevertheless made no difference and so in a sense was not important after all. This second intuition can now be represented by constructing the appropriate *relative* causal efficacy – what is the strength of Holmes's shot relative to that of Watson? Our formula for this, recall, is:

$$E(C \& W_1 \& D_0) - E(C_0 \& W_2 \& D)$$

The left-hand term again equals 1, since Moriarty dies if Holmes shoots. But now the right-hand term is 1 too, since if Watson were to shoot then Moriarty would still die even in the absence of Holmes's shot. Accordingly, the efficacy of Holmes's shot now = 1 – 1 = 0. In words, if Watson was going to shoot anyway then Holmes's shot indeed makes no difference.

We had two seemingly incompatible intuitions in the Holmes–Moriarty example. I conclude that the resolution is that one of those intuitions corresponds to the *absolute* efficacy of Holmes's shot and the other to its *relative* efficacy. Alternatively put, one corresponds to a counterfactual of $C_0 \& D_0$ and the other to a counterfactual of C_0 & D.[11]

[11] In an analogous context, Sober (1988: 317–318) comments: 'I therefore seem to find myself in the paradoxical position of saying that genes can be a cause of height, even if they are judged to have zero magnitude … causes may make no difference, but they are causes nonetheless… [But] perhaps this air of paradox can be dispelled … it is not hard to fathom how causes can fail to be necessary for their effects.' He then points out that Holmes's shot was the cause of Moriarty's death even though Watson's plans meant that Holmes's shot made no difference. But I think the 'air of paradox' here is better explained as being just a conflation of absolute and relative causal efficacies. Thus Holmes's shot is a cause but if Watson would have shot anyway then it has zero *relative* efficacy (relative, that is, to the case of Watson shooting instead). However, there would only be a real paradox if – as is not the case – the *absolute* efficacy of Holmes's shot were also assigned to be zero. Once we speak only of absolute efficacies then we again seem to track general causation obediently.

Note though that this analysis does trade in turn on the particular counterfactual term we inserted into the formula for absolute efficacy. In particular, in that term it is 'held fixed' that Watson did not shoot. Arguably this does capture what comes naturally to our intuition when considering the efficacy of Holmes's shot, but nonetheless here we are touching on deeper issues since in the wider philosophical literature such an interpretation of the counterfactual 'Holmes did not shoot' would certainly be non-standard. Lewis's approach to evaluating counterfactuals, for instance, would normally be thought to imply in this context that the counterfactual of 'Holmes did not shoot' should include that Watson does shoot, and hence would endorse a zero efficacy result. Indeed for just this reason Lewis had to set up his original (1973) counterfactual definition of causation very carefully so as still to be able to endorse Holmes's shot as a cause in cases such as this (i.e. cases of so-called 'early pre-emption'). An alternative approach to the evaluation of counterfactuals is provided by the more recent literature on causal modelling, although the particular issue of intuitions in pre-emption cases remains problematic even there. This is not the place to pursue a topic long notorious in the causation literature generally. But a fuller discussion of the evaluation of counterfactuals in biology does follow in the final section.

Physics and biology revisited

Recall our example from physics of a Newtonian particle acted upon by both gravity and electricity. What is the causal efficacy here of, say, gravity? For C_1 = the actual level of gravity, and C_0 = a neutral (i.e. zero) level of gravity, our definition yields:

$$E(C_1 \& W_1) - E(C_0 \& W_0) = (\text{the particle's motion with gravity}) - (\text{the particle's motion with no gravity})$$

The way the example was introduced, the *relative* efficacy of gravity is exactly the same as its absolute one. However, the two could have diverged if we had adopted a different choice of counterfactual. Suppose we were comparing the strength of gravity on Earth with that on the Moon. Then C_1 would be the Earth's gravity as before but C_2 would be some lesser but now non-zero alternative level of gravity, corresponding to its strength on the Moon. Now the calculation would run:

$$\begin{aligned}\text{Efficacy of } C_1 \text{ relative to } C_2 &= E(C_1 \& W_1) - E(C_2 \& W_2)\\ &= (\text{the particle's motion with Earth's gravity})\\ &\quad - (\text{the particle's motion with Moon's gravity})\\ &= [E(C_1 \& W_1) - E(C_0 \& W_0)] - [E(C_2 \& W_2)\\ &\quad - E(C_0 \& W_0)]\\ &= [\text{efficacy of Earth's gravity}]\\ &\quad - [\text{efficacy of Moon's gravity}]\end{aligned}$$

(Here, W_2 = the background conditions given C_2.)

There are two different questions here: 'how much difference does Earth's gravity make compared to the Moon's gravity?', and 'how much difference does Earth's gravity make compared to no gravity at all?' The latter is an absolute causal efficacy and the former a relative one. The difference between the questions is entirely down to choice of counterfactual – either Moon or zero. Often the implicit choice of counterfactual will in fact be zero either way, in which case the absolute and relative efficacies coincide and there will not be even the appearance of ambiguity.

This explains why the issue seemed so unproblematic in our Newtonian case and indeed in many everyday examples too. But in social science, for instance, the appropriate choice of counterfactual is often far less obvious. And as we saw, complications also arise in biology – so return now to our example of the genetic and environmental influences on corn plants. Suppose first we are concerned with the singleton case of an Expensive strain of plant treated with SuperGro fertiliser (i.e. the top-left cell in our original Table 1). What does our formula say here? For C_1 = SuperGro (say), and C_0 = the neutral (i.e. zero) level of SuperGro:

$$\text{Efficacy of } C_1 = E(C_1 \,\&\, W_1 - E(C_0 \,\&\, W_0)$$

The left-hand term is the plant's actual height of 145 cm. The right-hand term is whatever height the plant *would* have reached given zero addition of SuperGro. The difference between the two terms would then yield the efficacy (for the Expensive strain, in these circumstances) of SuperGro.

What of 'environment' more generally? Let C = all the non-genetic input. The neutral level C_0 of such a C would presumably be any environment, such as being marooned in interstellar space, in which the plant would not grow at all. For this choice of C_0 we find that, in the case (as in the top-left cell of Table 1) that actual environment = SuperGro (plus actual background conditions), and actual genes = Expensive strain:

$$\begin{aligned}\text{Efficacy of environment } C &= E(C \,\&\, W_1) - E(C_0 \,\&\, W_0)\\ &= (\text{plant's height with actual environment and genes})\\ &\quad - (\text{plant's height with actual genes but only the}\\ &\qquad \text{interstellar environmental input,}\\ &\qquad \text{i.e. so that the plant died immediately})\\ &= (\text{plant's actual height of 145 cm}) - 0 = 145 \text{ cm}\end{aligned}$$

Similarly for genes, let D = the plant's actual genetic input of the Expensive strain, and let D_0 = the neutral level of genetic input, in this case either zero genes at all or at any rate some genotype such that no plant height developed. Then (for appropriate W_2):

$$\begin{aligned}\text{Efficacy of genes } D &= E(D \,\&\, W_1) - E(D_0 \,\&\, W_2)\\ &= (\text{plant's height with actual environment and genes})\\ &\quad - (\text{plant's height with actual environment}\\ &\qquad \text{but no genetic input})\\ &= (\text{plant's actual height of 145 cm}) - 0 = 145 \text{ cm}\end{aligned}$$

Therefore genes and environment each have *equal* absolute causal efficacies of 145 cm; both have 'full potency'.[12]

Turn next to some *relative* efficacies, again calculated for the Expensive-SuperGro plant from the top-left cell of Table 1. Consider again C_1 =

[12] It might seem paradoxical that each of the two inputs could individually be awarded 'full' causal efficacy since this appears to imply that their efficacy together will be more than the total effect. So should they not, as it were, instead only have half each? But consider the two inputs' *joint* strength: if C = (actual genes & environment) and C_0 = the neutral levels of each, then that joint strength = $E(C \,\&\, W_1) - E(C_0 \,\&\, W_0)$ = (plant's actual height) – 0 = 145 cm again. So no causal efficacy is ever calculated to be greater than the total effect, and there is no paradox.

SuperGro fertiliser, and let C_2 = the average of the two alternative fertilisers, namely FastGro and Green.[13] Then (for appropriate W_2):

$$\text{Efficacy of SuperGro } C_1 \text{ relative to the FastGro/Green } C_2$$
$$= E(C_1 \,\&\, W_1) - E(C_2 \,\&\, W_2) = 145 - 0.5(115 + 85) = 45 \text{ cm}$$

Verbally, choosing SuperGro rather than the alternatives made (on average) a difference of 45 cm to the plant's height.

The analogous calculation for the Expensive strain of genes is:

$$\text{Efficacy of Expensive genes } D_1 \text{ relative to the Normal/Cheap genes } D_2$$
$$= E(D_1 \,\&\, W_1) - E(D_2 \,\&\, W_2) = 145 - 0.5(140 + 135) = 7.5 \text{ cm}$$

Verbally, choosing Expensive genes rather than the alternatives made on average a difference of only 7.5 cm to the plant's height. Therefore the environmental input's relative efficacy in this case is much larger than the genetic input's.

We saw earlier that there seems to be a duality in our understanding of causal efficacy in the biological context. On one understanding the environment is more important than genes (in the particular population of our example), while on the other comparing the two at all is meaningless. Our two different calculations above now capture this duality. The relative (to C_2 and D_2) efficacies, on the one hand, capture the sense in which varying the environment makes more difference than does varying the genes. The absolute efficacies (i.e. those relativised to C_0 and D_0), on the other hand, capture the sense in which the two inputs are equally and inseparably necessary to any plant height at all.

I conclude that therefore it is wrong to speak of there being two distinct senses of causal efficacy in biology. *Both* the sense in which genes and environment are entangled symmetrically, *and* the alternative sense in which one of them may after all be stronger than the other, can be adequately represented in terms of the same basic formula. They simply correspond to different choices of counterfactual.

ANOVA and counterfactuals: a critique

Recall again Table 1, showing fictional plant heights associated with various gene-environment combinations. Both our own formula and ANOVA agree

[13] As this calculation shows, our formula for relative efficacy is readily extendable to cases of more than one counterfactual, so long as we specify a suitable weighting across those counterfactuals. Indeed one weakness of ANOVA, in contrast, is precisely the difficulty of flexibly altering such weightings – instead we are always in effect constrained by the actual sample available. The entry in each cell in an ANOVA table is the average score for the subsample of treatments composed of that particular combination of inputs. Problems arise for full ANOVA inference if the sizes of the different subsamples vary. While ANOVA does offer a variety of methods for accommodating such asymmetric cases, the aim is always to allow again an equal weighting for each cell.

that environment is the more important causal factor in this example, but this apparent harmony conceals some fundamental difficulties. I shall discuss here only perhaps the most glaring of them, namely ANOVA's inflexibility with regard to choice of counterfactual.[14]

As we saw, classical ANOVA is applicable only to populations as a whole. Therefore for a proper comparison we need to be clear on how to apply our own definition of causal efficacy also to such *group* cases. Consider in our example the Expensive plant strain. For each Expensive plant, the (absolute) efficacy is just the height of that plant, as we saw. Therefore the group efficacy of Expensive – that is, the total extra plant height in that group attributable to the presence of the Expensive genotype compared to no genotype at all – is the number of individual Expensive plants in the group multiplied by their average height. (Equivalently, the group efficacy can be expressed as the sum of the individual efficacies.) Its value therefore depends both on the efficacy of Expensive on each individual plant, *and* on the ubiquity of Expensive plants in the population as a whole (see also Sober et al. (1992) on this point). It follows that the group-efficacy of 'Expensive' is, more strictly speaking, actually a group-efficacy of a particular *distribution* of Expensive. In our example, this distribution was one-third of the total population of plants, equally split between each fertiliser treatment – i.e. exactly one-third of the SuperGro-treated plants, and one-third of those treated with each of the other two fertilisers too. A different distribution, for instance if one-half of all plants had been Expensive and only one-quarter had been each of the other two strains, would have resulted in a different group efficacy.[15]

A similar analysis applies to *relative* group efficacies: again, an efficacy will be of one distribution of a cause (or causes) relative to a different distribution. In our example, the overall average plant height for the given distributions of plant strains and fertilisers was 110 cm. But suppose that, instead of an equal one-third split between the Expensive, Normal and Cheap plant strains (say, distribution 'A'), we made two-thirds of the plants Expensive and split the remaining one-third between Normal and Cheap equally. This new distribution (B) of plant strains, thus weighted now more towards

[14] A second weakness is the inappropriateness in this context of ANOVA's concentration on variance – rather than level – of effect, on which see Northcott (forthcoming) and also Lewontin (1974). Others include ANOVA's treatment of interaction effects (also not discussed here), and its restriction to group rather than singleton cases.

[15] In general (although not for the figures in our simple numerical example) we would also obtain a different group efficacy if Expensive were proportionally more common among those plants treated with some rather than other fertilisers, because of the possibility of interaction effects between choice of plant strain and choice of fertiliser. For simplicity I ignore this possibility here, but one advantage of our definition of causal efficacy is its ability to accommodate such interaction effects straightforwardly.

Expensive, would result in a new average plant height – say (as would follow from the values of Table 1), 112.5 cm rather than the previous 110 cm. We can imagine further a third distribution (C) of plant strains, this time one whereby *every* plant is Expensive, and – for the given distribution of fertilisers – the average plant height for this third population would be 115 cm. Thus these three different distributions of plant strain would yield average population heights of 110, 112.5 and 115 cm. (This corresponds to the first column of Table 3.)

We can next imagine a similar varying of the distribution of *fertilisers*. For the actual distribution of plant strains (A) we saw that the actual distribution of fertilisers (X) yields an average height of 110 cm. But suppose that instead of an equal one-third split, instead a proportion 0.4 of the plants were treated with SuperGro and 0.3 each with FastGro and Green (distribution Y), and suppose further that this new distribution would yield an average plant height of 113. And imagine a third distribution (Z) of fertilisers, this time split one-half SuperGro and one-quarter each of the other two, yielding an average height of 117.5. Using the values of Table 1, we can in addition calculate the counterfactual average heights that would result given each combination of these new fertiliser and plant strain distributions (again assuming no interaction effects) – these are given in Table 3.

Table 3. Fictional group data for plant heights and gene-environment combinations.

Fertiliser distributions (proportions SuperGro–FastGro–Green):		X (1/3 each)	Y (4-.3-.3)	Z (1/2-1/4-1/4)
Plant strain distributions (proportions Expensive–Normal–Cheap)	A (1/3 each)	110	113	117.5
	B (2/3-1/6-1/6)	112.5	115.5	120
	C (1-0-0)	115	118	122.5

So the actual population, with a one-third distribution across both plant strains and fertilisers and with an average height of 110 cm, is represented in the top-left cell (A-X). Entries across the top row and down the first column represent counterfactual populations obtained, respectively, by varying the input of one of environment or genes. The other entries represent counterfactual populations obtained by varying both.

It is now easy to see how the relative group-efficacy of a (distribution of a) cause is exactly analogous in form to that for an individual. For instance, what is the relative efficacy of genes in this example? As with the singleton case, it depends critically on which counterfactual we use for comparison (as well as on background conditions). Given fertiliser distribution X, changing from genetic distribution A to distribution B yields an increase in average height of 2.5 cm

(112.5 – 110), whereas changing from A to C yields one of 5 cm. Different relative group efficacies for the environment can be worked out similarly: along the top row, switching from X to Y yields an increased average height of 3 cm, from X to Z one of 7.5 cm, and from Y to Z one of 4.5 cm. Which of 'genes' or 'environment' has the greater relative efficacy therefore *varies* depending on which counterfactual (and background conditions) is chosen. There simply is no univocal answer.

Return now to ANOVA. In contrast to our formula, it offers no such flexibility. Rather, the causal efficacy of genes in the actual (A-X) population is calculated and then can only ever take this one value. There is no mechanism for it varying with choice of counterfactual; instead, classical ANOVA simply never even incorporates the notion of a counterfactual. Its calculations of causal efficacy are thus extremely inflexible – and unsatisfactorily so. (See also Lewontin (1974) for extended discussion of related complaints.)

For example, suppose that a farmer wanted to know how to increase her plant's average yield – for a fixed budget, should she target genes or environment? Surely the answer would be critically dependent on what the available alternative genetic and environmental distributions were. Starting from the actual A-X population, if (for a given budget) the available alternative genetic distribution were B whereas the alternative environmental distribution were Z, then clearly she should spend her money changing to the latter. This is because switching plant strains from A to B would only increase average plant height by 2.5 cm, whereas switching the fertiliser distribution from X to Z would increase it by 7.5. Suppose though, to vary the case, that the alternative genetic distribution were actually C and the alternative environmental one Y. Now switching genes would increase average plant yield by 5 cm whereas switching environment only increase it by 3, and so of course the farmer should now spend her money on the new plant strain instead of new fertiliser. Yet other distributions of plant strain or fertiliser would yield yet other recommendations. The key point is that there is not necessarily any *single* recommendation in favour of either genes or environment. Rather, all will depend on circumstance, i.e. on what are the feasible alternative genetic and environmental inputs, i.e. on choice of counterfactual.

But ANOVA could only ever inflexibly give a single recommendation one way or the other, regardless of which counterfactuals might be salient. Its initial calculation (for the actual A-X population) would adjudge one of genes or environment the stronger cause, and that would be the end of matter. Therefore, since as we have just seen *either* of genes or environment may have the higher relative efficacy depending on our choice of counterfactual, so for at least some cases ANOVA's judgment will inevitably be wrong.

The same applies to any intervention. In this example, it was just a farmer who would have been ill advised always to follow ANOVA's recommendation,

but similar remarks apply in broader social contexts to choice of *policy* intervention. The point is that ANOVA makes a one-off judgment based on the actual range of inputs that happen to obtain in the population now, whereas the crucial thing from an intervention point of view is the effect *relative* to the salient counterfactual alternative. That is, the crucial thing is the choice of counterfactual – precisely the issue that ANOVA ignores.[16]

Remedies and counterarguments

How might ANOVA be adjusted or defended as a definition of causal efficacy? I shall survey, but reject, some suggestions.

Direct incorporation of counterfactuals

Why not, as it were, graft onto the ANOVA procedure a provision for choice of counterfactual? The thought would be that, first we could perform an ANOVA to calculate the efficacies of genes and environment in the actual population (A-X), next likewise perform an ANOVA on the relevant *counterfactual* population (B-X, say), and then perhaps subtract the results of one from the other in order to see how much difference each of genes and environment make. That way, scores for causal efficacy would indeed vary with choice of counterfactual, just as I have been urging.

But it would be a mistake to think that the suggested method provides any real succour for ANOVA. First, the procedure is quite unmotivated within the ANOVA framework, so even suggesting it is therefore already a concession to our own approach. Second, it turns out in more complicated examples that the exact weightings to put on each cell in the counterfactual population can be quite an intricate calculation (Northcott forthcoming). It requires in particular both the controlled-experiment sensibility characteristic of the counterfactual but not ANOVA approach, and also the capacity easily to incorporate asymmetric weightings – which is a problem for ANOVA (footnote 13). Third, the suggestion in any case remedies just one element in a wide portfolio of difficulties (footnote 14). Thus even on the rosiest view it could represent only a very incomplete salvation.

[16] This also leads to another, more subtle kind of inflexibility regarding counterfactuals. Changing just one input may in principle alter *any* cell in the table via indirect effects. For instance, increasing the proportion of plants that are Expensive may also alter the results for the *other* plant strains, perhaps via a changed impact on the field's overall nutrient balance or some such. Epistemically, even when we do not know them for sure, we should be free to fill in at least our best guess as to what these indirect effects might be. Our own formula can accommodate this straightforwardly via the different background conditions W in each term. But ANOVA is committed to there being no such variation, i.e. to just a single fixed table of results, and hence has no way of incorporating indirect effects or even our best guesses about them.

ANOVA across the meta-population

The fundamental incompatibility of the ANOVA technique with an appropriate incorporation of counterfactuals, also lies at the heart of the failure of two further possible remedies. First, another response that has been suggested to me is to perform an ANOVA on Table 3. That is, in the face of this range of counterfactual possibilities could we not still perform an ANOVA across all of them too, that is across the counterfactual combinations of A, B and C, and X, Y and Z? We could then use such a 'meta-ANOVA' to advise us which of genes and environment has a greater impact on plant heights, at least in this particular 'population of populations'.

I shall criticise this proposal in a moment. Before that though, notice that whereas above we focused on why ANOVA is unsatisfactory when comparing different populations, now the issue would be the different one of why it is unsatisfactory even *within* a given population (of populations). We can note immediately that the putative meta-ANOVA would still yield only a univocal answer as to which of genes and environment was the more efficacious whereas we saw above that really this judgment should be sensitive to varying circumstance. It also turns out that this new case is equivalent – and has equivalent weaknesses – to a proposal to extend ANOVA to singleton cases, so I shall deal now with those two both together.

Extension to the singleton case

A familiar limitation of ANOVA is its applicability only to group rather than singleton cases, but consider a possible extension of the procedure that would remedy this. Take a single corn plant, of a particular strain and treated with a particular fertiliser. We can imagine various possible alternative strains and compare the plant's actual height with those heights it would have attained had it been one of those alternative strains. Likewise we can also imagine alternative fertilisers with which it might have been treated, again together with associated hypothetical heights. The proposal is that we can formulate a table of these various *counterfactual* possibilities, perform an ANOVA on that table, and then apply the rankings of genes and environment from that ANOVA to the actual singleton plant. In this way ANOVA could after all be applied to singleton cases, via a particular selection of counterfactual alternatives.[17]

[17] I see this as being precisely the suggestion put forward in Sober (1988: 309–312). An anonymous referee disputes that reading, judging instead that what Sober was in fact proposing is to fill out the slots in the table with *actual* data – thereby offering a way to evaluate singleton causal efficacies using only actual data, much as my own scheme also does (see later). While I believe that my reading here is the correct one, it is also true that the relevant passages are not entirely clear. Regardless of whether or not the position discussed really is that of Sober (1988) though, of course the critique of it in the text still stands.

Now return to the proposed meta-ANOVA. In our example, we began with the A-X population. Classically, we could perform an ANOVA on the data from that population (Table 1) in order to evaluate the causal efficacies of genes and environment. The alternative meta-ANOVA proposes that, having compiled the extended table ABC-XYZ of counterfactual populations (Table 3), we perform an ANOVA on *that* table and then interpret the results to be the causal efficacies of genes and environment in the original A-X population. This is therefore exactly analogous to the suggestion above for how to apply ANOVA to singleton cases. The only difference is that instead of calculating the efficacies for an individual plant by performing an ANOVA across a table of counterfactual individuals (of which the actual plant is only one cell), now we are calculating the efficacies for an individual *population* of plants by performing an ANOVA across a table of counterfactual *populations* (of which the actual population is only one cell).

But to see why neither proposal is satisfactory, consider a simple decision problem. A farmer is stuck with an old greenhouse and an old strain of plant. Her yield per plant at the moment is 4 units. Suppose that there exist two equally expensive improvements but that she can afford only one of them. Of course, she will choose the one that improves her yield the most, in other words the one with the greater causal efficacy. The first alternative is to replace her old greenhouse with a new one, which would improve the yield to 6. The second is to leave her greenhouse alone and instead to replace her old strain of plant with a genetically modified new strain. Doing the latter would improve her yield to 8. Which of the improvements should she spend her money on? The answer is obvious: on the new plant strain rather than on the new greenhouse.

Our definition of causal efficacy represents this reasoning successfully. Let C_2 = the new plant strain, C_1 = the old one, D_2 = the new greenhouse, and D_1 = the old one. And let different W_i be the various appropriately specified background conditions. Then, in the farmer's initial circumstances:

(1) Efficacy of the new plant strain C_2 relative to the old plant strain C_1
$= E(C_2 \,\&\, W_2 \,\&\, D_1) - E(C_1 \,\&\, W_1 \,\&\, D_1)$
$= 8 - 4 = 4$
(2) Efficacy of the new greenhouse D_2 relative to the old greenhouse D_1
$= E(D_1 \,\&\, W_3 \,\&\, C_1) - E(D_2 \,\&\, W_4 \,\&\, C_1)$
$= 6 - 4 = 2$

Therefore genes make the more difference in this case, i.e. have the higher relative efficacy, and the farmer is correctly recommended to invest in the new plant strain rather than in the new greenhouse.

So far, so straightforward. But suppose that the full table of yields is as in Table 4.

Our own calculations remain as stated above. But now consider what ANOVA says. The sum of squares across the genetic MAs is zero, whereas that across the greenhouse MAs is: $2[(6-5)^2 + (4-5)^2] = 4$. Therefore ANOVA is

Table 4. Fictional data for singleton gene-environment choice.

	Old greenhouse	New greenhouse	MA
Genes old	4	6	5
Genes new	8	2	5
MA	6	4	Grand mean: 5

forced to conclude that it is the environmental input that is the more efficacious cause here and so must actually advise the farmer to upgrade the greenhouse rather than the plant strain. What has gone wrong?

The problem is of course the very low yield of 2 scored by the combination of new genes and new greenhouse due to some strongly negative interaction effect. The key point is that this interaction effect should be *irrelevant* to the farmer's decision since by assumption she has enough money to change only one of her inputs. That is, we are concerned only with how much difference, compared to the original set-up, each of the new inputs makes individually. When defining the impact of the new greenhouse we of course did this while holding constant the other factor, i.e. the plant breed. This again is just the logic of controlled experiment – when assessing the impact of a given cause, one tries to hold constant all other causally relevant factors. Similarly here, when calculating the impact of switching to a new greenhouse the one thing we surely want to *avoid* is allowing the plant breed to vary too. But in effect that is just what ANOVA does. By focusing on the sums of squares across the MAs it necessarily incorporates information from the irrelevant bottom-right-hand cell, leading in this case to the perverse pro-greenhouse advice.[18]

It is important always to keep clearly in mind exactly which causal efficacy we are concerned with. For instance, we could imagine a whole population of farmers, some with the new greenhouse some with the old, some with the new plant strain some with the old, some with both, some with neither. Across this whole population, assessment of the average impact of genes and environment would indeed need to take into account the negative interaction effect in the bottom-right-hand corner. But the causal efficacy across a whole population would be a *group* efficacy (see previously) and not the singleton one at issue here. Likewise, if we were concerned with the singleton problem starting from the top-right corner then when calculating the causal efficacy for genes we would indeed be interested in the change down to the bottom-right corner. And similarly the efficacy of environment starting from the bottom-left corner

[18] In reality, of course, presumably no one would follow ANOVA so blindly as actually to choose the new greenhouse here, so are we attacking merely a straw man? After all, biologists intelligently manipulate their choice and design of experiments, and apply common sense to avoid obviously absurd conclusions. But the telling point is the very need for such manipulations in the first place. If in practice we are forced to trim away from strict adherence to ANOVA then it surely cannot be a satisfactory definition of causal efficacy. For this reason perhaps, often ANOVA is not cited as such explicitly. Yet as we saw (Table 2), implicitly it often is surely so *used*.

would also want to take account of the bottom-right corner. Or perhaps we might start from the top-left corner and be interested in the joint efficacy of adding both the new plant breed and new greenhouse simultaneously. In all these cases, the bottom-right corner would be relevant. But none of these cases is our case. And for our case, the bottom-right corner is *not* relevant.

Summing up the preceding sections, ANOVA applied to singleton cases pays heed to factors that are not relevant, and when applied to group cases pays no heed to factors that *are* relevant. So in neither case does it calculate causal efficacies satisfactorily.

Evaluating counterfactuals

Counterfactuals and actual data

The argument in this paper is that what we need in biology is a flexible, context-dependent concept of causal efficacy based on comparison of the actual level of effect with a counterfactual one. On my account indeed, the very notion of causal efficacy makes implicit contrastive appeal to counterfactuals. However, even if all this is granted, still we must address the classic philosophical issue of just how the relevant counterfactuals are to be evaluated.[19]

Obviously, by definition no direct measurement of counterfactuals is possible, so as a matter of scientific practice we are forced to seek some *proxy* for them from actual data. In an experimental setting we do this by controlling as best we can for all other causal factors and then seeing how much change in an effect term we get when changing the cause of interest. In this way we can get an estimate of the causal efficacy entirely from actual data. As noted, our formula can be seen as applying the logic underlying this method to the definition of causal efficacies generally.

Such a proposal has much in common with the recent literature in causal modelling (Spirtes et al. 2000; Pearl 2000; Woodward 2003). The understanding there of causal efficacy, and indeed of causation itself, is in terms of a kind of contextualised counterfactual dependence – what *would* be the impact of changing the value of this variable in these particular circumstances? The formal apparatus of causally interpreted directed acyclic graphs is then developed so that the relevant counterfactuals may be evaluated via reference to an *intervention* on those graphs. Such an intervention in turn represents the changing of the level of one input. Typically it is assumed that such an intervention in itself leaves the causal relations in a graph unaffected. This enables the impact on any other variable of the change of the level of one input to be traced formally as the impact of an intervention. The relevance to us is that by evaluating counterfactuals with reference to interventions in graphs in this way, the requisite causal sensibility is automatically

[19] I am grateful to an anonymous referee for pressing me on the importance of this.

built in, so to speak. The key is to find a real-world causal system that will indeed remain structurally stable when we change the value of one of the inputs. In such a system – but only in such a system – we can then read off a causal efficacy directly as being the change in an effect variable that follows the change in a cause variable. The kindredness to the method of controlled experiment is obvious.

Of course, deciding exactly which real-world systems meet the appropriate stability requirement will likely be based on some complex function of data and background knowledge. The crucial point is that not just any actual data is suitable, and in the case even of appropriate data not just any way of analysing it is suitable. Rather, the notion of a causal system stable under the relevant intervention represents a normative *ideal* that governs our evaluations of counterfactuals and hence in turn our inferences of causal efficacies. In practice the choice of the appropriate data may often be natural and obvious, arguably as when analysing the various examples in this paper, for instance.

The above approach to evaluating counterfactuals is distinct from Lewis's familiar apparatus of possible worlds, which appeals instead to a similarity metric based on nomological considerations. Nevertheless, depending on how exactly in any one case similarity is understood, of course it is possible to interpret a Lewisian scheme too as endorsing controlled-experiment methods for evaluating counterfactuals (at least usually – on which see Woodward 2003). Thus in practice when it comes to evaluating counterfactuals by means of actual data, the same procedure may be endorsed.[20]

ANOVA and data

Sometimes the right data might be hard to acquire. But there are two distinct issues at play here, conceptual and epistemological, and it is not clear why a more difficult epistemological situation should make any difference to the conceptual question. Hence there is no reason why our *definition* of causal efficacy should change just because of data difficulties.

Perhaps it might still be thought that ANOVA could be a source of useful information for evaluating the relevant counterfactuals. But I say that whereas the data on which ANOVA works may indeed be useful for that, the ANOVA procedure itself is not. Maybe, in certain difficult epistemological situations, ANOVA could even be a good substitute for our counterfactual definition? But

[20] It may sometimes be that a particular counterfactual is especially vague or indeterminate. In such a case, I agree with Sober that 'to the degree this is so … the question of causal magnitude also is vague and indeterminate' (1988: 310). In the context of evaluation via a causal model, either it would be unclear which particular causal model is the appropriate one, or else unclear which is the salient intervention within that model. In Lewis's scheme, likely the interpretation of the counterfactual's antecedent would be unclear. Either way, the problem for our purposes is that it would consequently in turn be left unclear exactly which, if any, proxy actual data is appropriate to evaluating the causal efficacy.

the argument of this paper has been carefully that it is not, or more precisely that it is not a good substitute for the operationalisation of that definition in terms of actual data.

Turn now to broader questions surrounding data. In practice, procedures of data collection are often designed precisely in accordance with our normative ideal. For instance, field trials of corn plants typically take great pains to have many plants growing under each treatment combination (so as to minimise the statistical impact of idiosyncratic one-off events such as a plant becoming infested), to avoid having choice of fertiliser correlate with being on the side of the field nearer a river or any other possibly causally relevant factor, and so on. That is, every care is taken to ensure that background conditions are kept as constant as possible and hence that actual data are as good a proxy as possible for the relevant counterfactual quantities.[21] In the notation of our formula, that means keeping as similar as possible W_1 and W_0 (except in so far as the change between C_0 and C may itself alter W).

Like other measures in statistics, such as Pearson's correlation coefficient, ANOVA reflects its positivist origins in being defined only over actual data. But on the approach advocated here our best *estimates* of causal efficacy are also defined over actual data, indeed often exactly the same data as ANOVA uses.[22] For, as mentioned, when collected with appropriate care ANOVA's table of data is certainly suitable as raw material for the assessment of causal efficacies – just not in the way that ANOVA does it.

Sometimes it will not be possible to read the value for the relevant counterfactual directly off the actual table of data but that value will nonetheless be deemed obvious in any case, for instance that with zero oxygen and water a plant would not have grown at all. One might speculate that ultimately the reason the value of this counterfactual seems obvious is precisely that our

[21] On our definition, to repeat the value of a causal efficacy depends not only on the actual level of effect but also on the counterfactual level that would have obtained under particular alternative circumstances. As now explained, our proxy here for that latter term is a second *actual* level of effect. Therefore it can seem that, as Sober (1988: 318) argues, "even if causality is local, the magnitude of causality need not be." But the non-local actual term is only a proxy for a counterfactual one, and the latter arguably *would* be 'local' in the sense implied.

[22] Still, given that our definition necessarily involves the evaluation of counterfactuals whereas ANOVA by contrast is defined purely from actual data, is not ANOVA therefore at least preferable on epistemological grounds? But I think this assumes an unwarrantedly pessimistic view of our ability to evaluate the relevant counterfactuals. On the account presented here, the method of controlled experiment – surely the gold standard of science if anything is – itself assumes the ability to evaluate counterfactuals (see also Woodward 2003 on this point). So in this respect the procedures endorsed by our definition are no different from those underpinning the paradigm successes of physical science. Thus on pain of writing off the latter the objection is not compelling. Indeed, one might turn the argument around. For many causal efficacies, very little new data is needed for their calculation – often just two cells in a table, as we have seen. Even for group efficacies, in effect the data requirement is often less than a whole table, depending on exactly which group efficacy is being calculated. ANOVA by contrast requires the entire table of data and so is often actually much *more* demanding than our own definition. The issue of data requirement would then become another point *against* ANOVA.

background knowledge is in effect itself appealing to various natural experiments, e.g. that all plants in circumstances otherwise causally similar have been observed to die when starved of oxygen and water. So even if not explicit in an ANOVA table, the proxy actual data – now located, as it were, in background knowledge – relevant to the evaluation of these other counterfactuals are still just as much informed by a controlled-experiment sensibility.

What if the data on which we conduct an ANOVA are taken from totally uncontrolled sources? For example, suppose we tracked crime rate data taken from different countries and constructed a table across different combinations of, say, liberality of penal code and numerical strength of policing. But suppose also that all the while we were taking no account of other factors likely causally relevant, e.g. local demographics regarding the number of young males, socioeconomic conditions, gun laws, and so on. I am sceptical that such data could provide useful information about causal efficacies, precisely because of (in our notation) the wildly varying background conditions W. (Northcott (forthcoming) demonstrates this point in rather more detail.) Rather than trying to draw conclusions from such uncontrolled data, our best bet would be to look for *other* data that better controlled for confounds, or perhaps we might disaggregate in suitable ways the existing data. The point is that in these circumstances ANOVA would offer no guidance for, nor even recommendation of, such essential procedures.

Conclusion

The very notion of causal efficacy appeals to counterfactuals and consequently represents a conceptual ideal to which we mortals stuck with actual data can merely aspire. But this is no barren ideal. Rather, it is one that informs just how we should collect, and then analyse, actual data. In particular, it shows us both why ANOVA is a bad way of analysing actual data and also shows us a better alternative way of analysing that very *same* data. Therefore the critique does not rest simply on a smug appeal to counterfactuals while, as it were, unfairly marooning ANOVA on an island of actuality.

General scientific practice reflects its agreement with our ideal via its emphasis on controls and experiments, and moreover by its successes amply demonstrates that ideal's viability. Thus exaggerated epistemological scruples should not sway us away from the counterfactual definition of causal efficacy in biology any more than in the rest of science. Rather, the question becomes only how best to operationalise it.

Acknowledgements

I would like to thank Nancy Cartwright, Elliott Sober and an anonymous referee for helpful discussion. Material from this paper also benefited from

being presented to the following audiences, to whom I am also grateful: the 'Causality: Metaphysics and Methods' research project at the London School of Economics; the Science Studies program at the University of California, San Diego; and a graduate student philosophy conference also at the University of California, San Diego.

References

Bliss C. 1967. Statistics in Biology. McGraw-Hill, New York.
Good I.J. 1961. A Causal Calculus parts I and II. British J. Phil. Sci. 11: 305–318 and 12, 43–51.
Humphreys P. 1989. The Chances of Explanation. Princeton University Press, Princeton, NJ.
Lewis D. 1973. Causation. J. Phil. 70: 556–567.
Lewontin R. 1974. Analysis of variance and analysis of causes. Am. J. Human Genet., 400–411.
Miller R. 1987. Fact Method. Princeton University Press, Princeton, NJ.
Northcott R. forthcoming, Pearson's wrong turning: against statistical measures of causal efficacy, Phil. Sci. Proc. 2004.
Pearl J. 2000. Causality. Cambridge University Press, New York.
Remington R.D. and Schork M.A. 1970. Statistics with Applications to the Biological and Health Sciences. Prentice-Hall, Englewood Cliffs, NJ.
Sober E. 1988. Apportioning Causal Responsibility. J. Phil. 85: 303–318.
Sober E., Wright E.O. and Levine A. 1992. Causal Asymmetries, In: Reconstructing Marxism. Verso, New York, pp. 129–175.
Sokal R.R. and Rohlf F.J. 1981, 1995. Biometry: The Principles and Practice of Statistics in Biological Research, 2nd and 3rd edns. Freeman, New York.
Spirtes P., Glymour C. and Scheines R. 2000. Causation, Prediction, and Search, 2nd edn. MIT Press, Cambridge, MA.
Woodward J. 2003. Making Things Happen: A Theory of Causal Explanation. Oxford University Press, New York.