



## CAN ANOVA MEASURE CAUSAL STRENGTH?

ROBERT NORTHCOTT

*Department of Philosophy, University of Missouri-St Louis  
St Louis, Missouri 63121-4400 USA*

E-MAIL: NORTHCOTTR@UMSL.EDU

### KEYWORDS

absolute, analysis of variance, causal strength, causation, data, group, individual, intervention, level, measurement, relative, variance

### ABSTRACT

*The statistical technique of analysis of variance is often used by biologists as a measure of the relative strength or importance of causal factors. I argue that it is a tool ill-suited to this purpose, on several grounds. I suggest a superior alternative and outline its implications. I finish with a diagnosis of the source of error—an unwitting inheritance of bad philosophy that now requires the remedy of better philosophy.*

### INTRODUCTION

**M**ORE THAN 30 YEARS AGO, Richard Lewontin wrote a seminal article (1974) critiquing the use of analysis of variance (hereafter “ANOVA”) to license interventions. In this paper, I reinforce and extend this critique by embedding it in contemporary philosophical analysis of causation. Lewontin’s misgivings about interventions are merely part of a range of concerns that then emerge. My specific target will be the widespread use of ANOVA in biology as a tool for measuring the relative strength of causal factors (or, equivalently, their “importance”). I shall propose a better-founded—and practicable—alternative conception of causal strength.

An underlying theme will be that philosophy can help scientific practice, and also that the misuse of ANOVA is itself a symp-

tom of philosophical ideas now widely accepted (in philosophy of science) to be anachronistic. The benefit of being explicit about philosophical underpinnings is, therefore, the freeing of biological practice from this unwitting inheritance.

### ANOVA AND CAUSAL STRENGTH

I shall assume a certain familiarity with the ANOVA technique. Nevertheless, a specific—albeit fictional—numerical illustration will be helpful (Table 1). Imagine plotting plant height (the dependent variable) against two independent variables, namely choice of fertilizer (Expensive-Average-Cheap) and choice of plant breed (GM-Normal-Old).

Which factor has more influence on plant height? Intuitively, in this population, plant height varies far more with fer-

**TABLE 1**  
*Fictional data for plant heights and fertilizer-breed combinations*

	Expensive	Average	Cheap	MA
GM Breed	85	55	25	<b>55</b>
Normal Breed	80	50	20	<b>50</b>
Old Breed	75	45	15	<b>45</b>
<b>MA</b>	<b>80</b>	<b>50</b>	<b>20</b>	<b>Grand</b>
				<b>Mean: 50</b>

\*“MA”= marginal average. For simplicity, I have assumed that there are no nonadditive interactive effects, that each cell has equal weighting, and that we are considering only two independent variables. Nothing of importance will turn on these assumptions.

tilizer than with breed. ANOVA captures this via the following (simplified) calculation, where “SS” = sum of squares:

$$\text{Breed SS} = 3[(55 - 50)^2 + (50 - 50)^2 + (45 - 50)^2] = 3(25 + 0 + 25) = 150$$

$$\text{Fertilizer SS} = 3[(80 - 50)^2 + (50 - 50)^2 + (20 - 50)^2] = 3(900 + 0 + 900) = 5400.$$

Because 5400 > 150, ANOVA declares that choice of fertilizer “explains” more of the height variation than does choice of breed. I am not concerned here with statistical inference, hence no reference to error terms.

Of course, the analysis of variance is used in biology for more than one purpose. But one such use is as a de facto measure of degree of causal impact of independent variables on dependent ones. For instance, in the example above ANOVA would in effect be used to license the claim that, in this population, choice of fertilizer is a stronger cause of plant height than is choice of breed.

It is worth emphasizing this latter claim about usage: ANOVA’s 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, it can hardly not be) interpreted as one of causal strength—i.e., of how much impact the cause variable has on the effect variable.

To make the claim seem plausible within the confines of this paper, consider the following actual subjects of ANOVA study (Table 2). These are studies chosen by standard textbooks precisely to illustrate

**TABLE 2**  
*Topics of actual 2-factor ANOVAs*

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 of potato tuber	degree of aphid infestation
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 fertilizer	quantity of phosphoric acid fertilizer	yield of corn
sodium concentration	potassium concentration	weight of tobacco leaves

This table is taken from Northcott 2006. The individual references for the data are contained in Sokal et al 1981, 1995; Bliss 1967.

appropriate usage. I take them as thus to be cases of reputable work and not unrepresentatively sloppy.

In each case, the choice of variables clearly suggests particular cause and effect relations, and so any results linking those variables quantitatively should indeed be interpreted as (claimed) causal strengths. As well as being widespread, albeit often only implicit, I shall now argue that this use of ANOVA is also mistaken. In order to explain why, let us first take a brief foray into philosophy.

#### A DEFINITION FROM PHILOSOPHY

How important is a particular cause of an effect? To know that, we must compare the level of effect given the presence of the cause with what the level of effect would have been without that cause. Generalized and in symbols, define the strength of a cause  $C$  with respect to an effect  $E$  to be:

$$E(C_1 \ \& \ W) - E(C_0 \ \& \ W). \quad (D)$$

$C_1$  is the actual level of  $C$ ,  $C_0$  is a baseline counterfactual level of  $C$ , and  $W$  represents background conditions.

In the simplest case,  $C_0$  will just be the absence of  $C$ . But often we are interested in the impact of a cause relative to some specific nonzero alternative, as I shall illustrate below. Moreover, the "absence" of a cause may sometimes make little sense anyway, as with temperature. Thus, in its right-hand term (D) cites the general formulation  $C_0$  rather than "absence of  $C$ ."

(D) is quite intuitive. For example, if  $E$  = the height of a plant, then we might compare  $E$  given the use of a fertilizer to  $E$  given no fertilizer in order to define the causal strength of that fertilizer. We could define the causal strength of extra watering similarly, in order to compare which of the two causes—fertilizer or extra watering—was the predominant cause of the plant's height. The formula also enables us to define easily the impact of using one fertilizer rather than another, simply by setting  $C_0$  = the other fertilizer (rather than no fertilizer at all). Notice also the importance of the background conditions  $W$ . For example, if  $W$  = a poor level of soil

nutrients, then the impact of adding fertilizer may be low. If  $W$  = a rich level of soil nutrients, then, by contrast, the impact may be high.

A key aspect of (D) is that it captures a controlled experiment sensibility. We want to compare the level of effect with and without the cause while *keeping all else equal*. For instance, it would be futile to compare the height of a plant with and without fertilizer if simultaneously we also switched to a more nutritious soil, because, obviously, that would now yield only the combined impacts of the two changes. For this reason, in (D) the background conditions  $W$  must be constant across the two terms. Strictly speaking, the background conditions *do* in fact vary across the formula because, as well as impacting  $E$ , the switch from  $C_0$  to  $C$  will also change  $W$ . But for our purposes we may ignore that technical wrinkle here.

The formulation of (D) reflects the common emphasis in the philosophical literature on causation's difference-making aspect. A cause is something that makes a difference to its effect. Thus, naturally, the strength of a cause is how *much* difference it makes. (For some philosophical details, see Lewis 1973, Wright et al 1992, Pearl 2000, Spirtes et al 2000, Woodward 2003, Northcott 2006. For a recent introduction to the ideas on causation mentioned in this paper, geared specifically to biologists, see Shipley 2000.) Much philosophical debate has focused on which—if any—formulation of difference-making might be an acceptable definition of causation. Other discussions have focused on exactly how counterfactuals, such as the right-hand term in (D), should be understood. Fortunately, we need not worry about either of these issues here. First, in biology the putative causes of an effect are usually uncontroversial; our concern is only with which of these causes is stronger. Second, (D) can be seen as a normative ideal, serving to guide which actual data we should use as proxies for the counterfactual terms. Given the typical set-up of biological experiments, in practice, this is usually uncontroversial, as I shall illustrate below.

TABLE 3  
*Fictional group data for plant heights and fertilizer-breed combinations*

Fertilizer distributions (proportions expensive-average-cheap):		F1 (1/3 each)	F2 (0.4-0.3-0.3)	F3 (1/2-1/4-1/4)
Breed Distributions				
(Proportions GM-Normal-Old):	B1 (1/3 each)	50	53	57.5
	B2 (2/3-1/6-1/6)	52.5	55.5	60
	B3 (1-0-0)	55	58	62.5

It is obvious that the strength of any given cause depends on the effect we are interested in. For instance, fertilizer may be a powerful cause of plant growth, but a negative cause of human growth. In our notation, C's causal strength is E-specific. But notice also an important further corollary of the above discussion—namely that, even once E is specified, there is still no unique value for C's causal strength. Instead, for any given C-E pair, the associated causal strength depends also on W and  $C_0$ . Thus, the notion of causal strength is necessarily a highly relativized and context-specific one, a fact that will prove significant.

To illustrate the simplicity of these basic ideas, I shall close this section by applying (D) to our fictional example in Table 1. The switch from Average to Expensive fertilizer, on a GM plant, increases height from 55 to 85. Thus, Expensive fertilizer, relative to Average fertilizer, for this effect in these conditions, has a causal strength of 30 units. Similarly, the causal strength of GM relative to Normal breed is 5 units. Indeed, formula (D) might seem so obvious and unobjectionable as to be rather uninformative. Yet, ANOVA contravenes this formula in multiple ways. Let us now explore some of these contraventions and their deleterious consequences.

#### INTERVENTIONS

Superficially, in Table 1, ANOVA and definition (D) agree: in this population, fertilizer is a more important cause of plant height than is breed. But look more closely—begin by applying (D) to the group case. (This is necessary to render (D) directly comparable with ANOVA, given that

the latter only delivers causal strength verdicts for populations or groups as a whole.) In the group case, the “cause” is not a single application of fertilizer or single choice of plant breed, but rather the particular *distribution* of such applications or breeds in a given population. Analogously to the individual case, and for analogous reasons, the causal strength of such distributions depends on which alternative distributions we are comparing them to.

Illustrating via our fictional example, label the actual population's distributions of breeds and fertilizers respectively B1 and F1. Suppose the average heights for particular alternative populations are as in Table 3. The top-left slot, i.e., the average plant height in the B1-F1 population, we already know from Table 1 to be 50 units (usually, we would need to specify the distributions in more detail than in Table 3, but the lack of interaction effects here allows us to simplify). What causal strengths can we read from Table 3? Consider first the fertilizer. The causal strength of distribution F2 relative to distribution F1 is 3 units (e.g., 53 – 50 in the top row); that of F3 relative to F1 is 7.5 units. For breed, switching from B1 to B2 yields a score of 2.5 units (e.g., 52.5 – 50 in the first column), whereas switching from B1 to B3 yields a score of 5 units. The key point is that there is no univocal causal strength for either fertilizer or breed. Rather, the values vary depending on choice of contrast class, just as per (D). Accordingly, we cannot say which of the two factors has the greater causal strength in any absolute sense. Relative to some contrast populations, fertilizer is the more important factor; relative to others, breed is instead. We can also readily imagine how results would vary with different choices

of background conditions  $W$ . Yet ANOVA yielded just the one absolute result, as we saw earlier, inflexibly endorsing fertilizer as the stronger cause in our actual B1-F1 population.

Why should this inflexibility matter? One reason is that causation, and thus strength of causation too, is intimately linked to interventions (this is a central point in the philosophical literature). In our example, should a farmer buy new fertilizer or a new breed? The answer depends crucially on the salient contrast class. If, for a given budget, the farmer has a choice between upgrading from F1 to F3 or from B1 to B2, then clearly the former will have a bigger positive impact on plant height (7.5 rather than 2.5), i.e., investing in new fertilizer rather than breed is preferred. But if the choice is either F1 to F2 or B1 to B3, a change of breed rather than fertilizer would have the greater impact (5 rather than 3). Overall, the farmer's intervention should be guided by which cause has the greater strength, breed or fertilizer, and this, in turn, clearly depends on the salient contrast class, just as captured by (D)—but not by ANOVA. ANOVA's results therefore do *not* give reliable advice regarding interventions.

Lewontin (1974) also emphasizes this conclusion, or, more particularly, that the findings of ANOVA cannot be extrapolated beyond the generating sample. In fact, they are also unreliable even *within* that sample, as I shall discuss in the coming sections. Lewontin appeals to norms of reaction as a better guide to interventions, as these will yield the quantity of effect for different values of a given input. Because they are calculated for fixed background conditions  $W$ , these reaction functions in effect replicate the controlled experiment procedure endorsed by (D). More generally, knowledge of norms of reaction enables a whole range of causal strengths to be calculated immediately. Notice though that any given norm of reaction will only hold for one particular set of background conditions  $W$ , and, therefore, will not necessarily be a reliable guide to interventions

if  $W$  alters (e.g., due to changed levels of oxygen or acidity).

Notice also that ANOVA is unable to address the intervention issue via its own incorporation of a comparative approach. For instance, referring back to Table 3, it might be thought that by comparing ANOVA's results for causal strength from the B1-F1 population with its results from the B2-F1 population, we would thereby get a guide to the impact of substituting breed distribution B2 for breed distribution B1. But, as well as being quite unmotivated within the ANOVA procedure, this maneuver simply does not work. In more complicated examples, the weighting to place on each counterfactual term is an intricate calculation requiring (D)'s controlled experiment sensibility. Typically, this calculation also requires asymmetric weightings across the cells in the table, which is problematic for ANOVA.

In general (although not for the figures in our simple numerical example), we would also obtain different results for causal strength if, for instance, the GM breed was proportionally more common among those plants treated with some rather than other fertilizers, because of the possibility of nonadditive breed-fertilizer interaction effects. This would count as a further strike for (D) over ANOVA, although, for simplicity, I ignore the possibility here. ANOVA's systematically poor handling of nonadditive interactions is a central theme of Lewontin (1974). (D), by contrast, deals with them naturally and easily (Northcott 2006).

#### GROUP VERSUS INDIVIDUAL

Often we may be interested in single causal strengths, such as those associated with an individual plant rather than with a whole population. As noted already, (D) can capture individual or group causal strengths according to interest, whereas ANOVA can yield results only at the population level. This has interesting implications for the truism that it makes no sense to apportion causal responsibility between genes and environment for an *individual's* phenotypic trait, given the continuous inter-



action of the two causal factors. Notice that, notwithstanding its orthodoxy, there is good reason to doubt the truth of this truism.

Consider an individual plant from the top-left cell of Table 1. Quantitative degrees of causal responsibility for the 85-unit height of this plant can sensibly be assigned to fertilizer and to breed. The key is that there is no single canonical apportionment of such responsibility; rather, all will vary with choice of contrast class. But then, as we have seen, *all* apportionments of causal responsibility so vary. It implies nothing unusual about the plant case in particular.

Simultaneously, we can make sense of the orthodox view—namely that, since genes and environment, crudely understood, are both clearly necessary for any plant height at all, it makes no sense to assign either factor more responsibility than the other. I think this stems from an implicit selection of contrast classes of absence. That is, we imagine a plant “without genes” (i.e., no plant at all) and “without environment” (i.e., no oxygen or water, perhaps), and, whatever those phrases mean, they presumably imply no plant growth and, hence, zero height. Thus, relative to these contrast classes, (D) tells us that both genes and environment have causal strengths of: (the plant’s actual height) – (zero height) = the plant’s actual height. In this case, the causal strengths of genes and environment are entirely symmetric. But in other cases, with different choices of contrast class and background conditions, this will not be so. Therefore, in Table 1, it may make perfect sense to declare a top-left cell plant’s genotype (i.e., GM breed) to be more responsible for its height than is its environment (i.e., Expensive fertilizer).

ANOVA represents one way of recognizing the sense in which genes and environment can be said to make asymmetric contributions to a trait. But a philosophically well-grounded account of causal strength likewise enables us to separate, where appropriate, the contributions of genes and environment. And this can be done in the case of individuals just as much as in the case of groups. (See Northcott 2006 for

fuller discussion of the issues in this section.)

#### RELATIVE VERSUS ABSOLUTE

So far, we have been assuming an absolute conception of causal strength, but there is an alternative, relative conception. We may illustrate this distinction by reference to a medical example. How important a cause of heart attacks is a bad diet? According to the relative understanding, this question should be interpreted as asking: Does diet have a greater impact than other causes of heart attacks? Such an understanding is associated with significance tests. These must first measure what proportion of total “noise,” so to speak, a factor is responsible for before then deciding whether such a proportion can reasonably be assigned to mere chance. When based upon the absolute understanding, by contrast, the question should be interpreted as asking: How important a cause is diet in its own right? This second understanding is associated with Galilean idealization. We are concerned with isolating the impact of one factor alone, having abstracted it away from background noise.

(D) is a measure of absolute causal strength, but is easily adjusted to capture the relative version, as follows:

$$[E(C_1 \& W) - E(C_0 \& W)] / E(C_1 \& W).$$

This yields, as desired, a factor’s causal strength relativized to the total level of effect. ANOVA, by contrast, is inflexibly committed *only* to the relative conception. By definition, it evaluates the variance due to one independent variable relative to the total variance of the dependent variable. The issue is essentially one of inflexibility. Looking at Table 2, although we might sometimes be interested in relative causal strength, it is obvious that often we are not. So, in all the latter cases, use of ANOVA is fundamentally inappropriate (Northcott 2005).

#### VARIANCE VERSUS LEVELS

Suppose I were interested not in maximizing average plant height, but instead in minimizing the *variance* of plant height.

Perhaps I am supplying a firm that, for some reason, greatly values product uniformity. Turning again to Table 2, it seems rare that this will be the focus of our interest; nevertheless, perhaps it might be in this case. (D) is able to reflect such an interest straightforwardly in this instance, simply by defining the effect of variable E to be the variance of plant height rather than the height itself. (As Massimo Pigliucci suggested to me, perhaps this would be useful when measuring the strengths of causes of phenotypic plasticity, if the latter is defined as across-environment variance.) Note that, variance being defined only with respect to populations rather than individuals, this version of (D) adjusted to a focus on variances will no longer be applicable to individual cases. It is no coincidence that ANOVA, being precommitted to variances rather than levels, is only ever applicable to group rather than singleton cases.

The real problem with ANOVA here is again its inflexibility. In particular, it is committed to tracking causal strength with respect to variance rather than level of effect. This follows from its definition. To illustrate, imagine Table 1 adjusted so that the variance of the fertilizers' marginal averages was reduced, even while keeping their mean unchanged. This would result in ANOVA granting fertilizer a *lower* score for causal strength than before, even though the average impact of fertilizer on plant height was unchanged. The reason is that ANOVA is sensitive to variance—not level—of effect. This would be fine if variance was what we were interested in. But, again, it is obvious just from looking at Table 2 of actual ANOVA applications, that often we are interested in level rather than variance. Accordingly, in all such cases ANOVA must be an inappropriate tool for measuring causal strength.

#### CONCLUSION: PHILOSOPHY BAD AND GOOD

The economist John Maynard Keynes famously remarked that "practical men, who believe themselves to be quite exempt from

any intellectual influence, are usually the slaves of some defunct economist" (1936: 383). Alas, sometimes defunct philosophers too. One hundred years ago, early positivism frowned on the notion of causation, seeing it as mystical metaphysics with no place in modern science. This distrust strongly influenced the formulation of statistical measures of strength of association such as ANOVA (Pearl 2000).

Nowadays, as we have seen, ANOVA is a standard tool in biology for measuring degree of causal impact of one variable upon another. But its anachronistically anti-causal origins have left it ill-suited to this latter purpose. Its form still reflects its inheritance, failing to incorporate the requisite controlled experiment sensibility, even though (at least in the domain of biology) philosophy has long since dropped that original hostility to causation. It is for this reason that ANOVA is formulated purely in terms of actual data, and essentially just tracks patterns of correlation within that data. Similar criticisms apply to related statistical measures, such as  $r^2$ , heritability, and others—all defined, broadly speaking, in terms of ratios of variances of actual data (Northcott 2005; Pigliucci and Kaplan 2006). Such statistical techniques are consequently all unsatisfactory measures of causal strength. Some other techniques in common use, such as odds and risk ratios, and randomized trials, are better.

But now more up-to-date philosophy can repay the debt by showing us that there is no substitute for a flexible, context-sensitive concept of causal strength. This requires comparison of the actual level of effect with a counterfactual one informed by the appropriate controlled experiment sensibility. Variables C and E must be treated asymmetrically, reflecting the asymmetry of cause and effect. Although ANOVA cannot provide this, we can easily define a formula that does, namely (D). Of course, by definition counterfactuals such as (D)'s right-hand term are not directly observable. As mentioned earlier, (D), therefore, serves as a *normative ideal* rather than a direct measure, guiding which proxy actual data we should

use instead. As we have seen, that, in turn, can prove an intricate, highly context-specific operation. ANOVA's crude one-size-fits-all formula cannot pass muster.

Ironically, often the relevant actual data will turn out to be drawn from exactly the same set that ANOVA operates on. For example, plant trials typically take care to control for other causes of height, such as quantity of water or soil acidity, and have large numbers of plants per plot so as to minimize the impact of occasional confounders such as infestation. The point is that (D) mandates better use than ANOVA of this very same data.

If, by contrast, data are not well controlled for confounds, then it may be that causal strengths cannot be reliably estimated from them. In such circumstances, our best bet is to look for other data that are better controlled, or perhaps we might disaggregate the existing data in suitable ways. Yet ANOVA would offer no guidance for, nor even recommendation of, such essential procedures.

Still, given that (D) necessarily involves the evaluation of counterfactuals whereas ANOVA, by contrast, is defined purely from actual data, is not ANOVA 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 in this paper, 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). In this respect, the procedures endorsed by (D) are no different from those underpinning the paradigm successes of physical science. Therefore, on pain of writing off the latter, the objection is not compelling. Indeed, one might turn the argument around. For many causal strengths,

very little new data are needed for their calculation—often just two cells in a table, as we have seen. Even for group strengths, in effect the data requirement is often less than a whole table, depending on exactly which group strength is being calculated. By contrast, ANOVA requires the entire table of data and so is often much more epistemologically demanding than (D). The issue of data requirement would then become another point against ANOVA.

It may be that many biologists already harbor some suspicions of ANOVA as a measure of causal strength. After all, when using it they often intelligently manipulate their choice and design of experiments, and apply common sense so as 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 stray away from strict adherence to ANOVA, then it cannot be a satisfactory definition of causal strength. For this reason perhaps, ANOVA is often not cited as such explicitly. Yet as we saw in Table 2, it is used as such implicitly.

In conclusion, I believe that science and philosophy can influence each other to mutual benefit. Indeed, many of the philosophical theories of causation appealed to in this paper have themselves been influenced by scientific practice. In return, philosophy can sometimes aid science. This is especially true of topics such as causation, whose explicit investigation has been much more the province of the former than of the latter. In particular, in the case of ANOVA and causal strength, let biology now be beholden to philosophy that is good rather than defunct.

#### ACKNOWLEDGMENTS

I would like to thank Massimo Pigliucci for helpful comments, and for organizing the conference at which this paper was presented.

#### REFERENCES

- Bliss C. 1967. *Statistics in Biology*. New York: McGraw-Hill.
- Lewis D. 1973. Causation. *Journal of Philosophy* 70(17): 556–67.
- Lewontin R. 1974. Analysis of variance and analysis of causes. *American Journal of Human Genetics* 26(30): 400–411.
- Maynard Keynes J. 1936. *The General Theory of Employment, Interest, and Money*. London (UK): Macmillan.
- Northcott R. 2005. Pearson's wrong turning: against



- statistical measures of causal efficacy. *Philosophy of Science* 72(5):900–912.
- Northcott R. 2006. Causal efficacy and the analysis of variance. *Biology and Philosophy* 21(2):253–276.
- Pearl J. 2000. *Causality: Models, Reasoning, and Inference*. New York: Cambridge University Press.
- Pigliucci M, Kaplan J. 2006. *Making Sense of Evolution: The Conceptual Foundations of Evolutionary Biology*. Chicago (IL): University of Chicago Press.
- Shiple B. 2000. *Cause and Correlation in Biology: A User's Guide to Path Analysis, Structural Equations, and Causal Inference*. Cambridge (UK) and New York: Cambridge University Press.
- Wright E O, Levine A, Sober E. 1992. Causal asymmetries. Pages 129–175 in *Reconstructing Marxism*. London (UK) and New York: Verso.
- Sokal R R, Rohlf F J. 1981. *Biometry: The Principles and Practice of Statistics in Biological Research*. Second Edition. New York: Freeman.
- Sokal R R, Rohlf F J. 1995. *Biometry: The Principles and Practice of Statistics in Biological Research*. Third Edition. New York: Freeman.
- Spirtes P, Glymour C, Scheines R. 2000. *Causation, Prediction, and Search*. Second Edition. Cambridge (MA): MIT Press.
- Woodward J. 2003. *Making Things Happen: A Theory of Causal Explanation*. New York: Oxford University Press.