Pearson's Wrong Turning: Against Statistical Measures of Causal Efficacy

Robert Northcott†‡

Standard statistical measures of strength of association, although pioneered by Pearson deliberately to be acausal, nowadays are routinely used to measure causal efficacy. But their acausal origins have left them ill suited to this latter purpose. I distinguish between two different conceptions of causal efficacy, and argue that: (1) Both conceptions can be useful; (2) The statistical measures only attempt to capture the first of them; (3) They are not fully successful even at this; (4) An alternative definition based more squarely on causal thinking not only captures the second conception, but also can capture the first one better too.

1. Introduction. Karl Pearson, one of the brilliant founders of modern statistics, was also a fervent opponent of the concept of causation, regarding it as unscientific metaphysical speculation. In the 1911 edition of his *The Grammar of Science*, he described it as "another fetish amidst the inscrutable arcana of even modern science" (vi), and elsewhere as a "fantasy" (122) and a "conceptual bondage" (165). His correlation coefficient r was from the start intended as a "measure of the intensity of association" (160), and thus in his eyes as a replacement for talk of causation. This is reflected in its definition:[1]

$$r^2 = [\mathrm{Cov}(XY)]^2/\mathrm{Var}(X) \cdot \mathrm{Var}(Y).$$

True to Pearson's positivism, it is defined purely in terms of actual data.

†To contact the author, please write to: Department of Philosophy, University of Missouri–St. Louis, 599 Lucas Hall (MC 73), One University Blvd., St. Louis, MO 63121-4499. e-mail: northcottr@umsl.edu.

‡I would like to thank Nancy Cartwright, and the audience at the Experimental Philosophy Laboratory at University of California, San Diego, for helpful comments on earlier versions of this paper.

1. It is common to summarize the strength of correlation between X and Y by citing r^2 rather than r, since this enables both positive and negative correlations to be compared directly.

Philosophy of Science, 72 (December 2005) pp. 900–912. 0031-8248/2005/7205-0021$10.00

Further, variables X and Y are treated entirely symmetrically and there is no appeal to causation anywhere.

Of course, actually causation did not disappear from scientific practice in the way that Pearson had anticipated. Moreover when, as is frequently the case nowadays, background knowledge suggests one of X and Y to be the cause of the other, Pearson's r is normally (indeed can hardly not be) interpreted as a measure of *causal efficacy*—how *much*, or how *strong*, an impact does the cause variable have on the effect variable? But this latterday embracing of causation is not reflected in r^2's formulation, which—faithful to Pearson's original anti-causal metaphysical commitment—remains unchanged. Similar remarks apply to the measure of causal efficacy yielded by Ronald Fisher's analysis of variance technique (ANOVA), which again is at heart a ratio of variances. For reasons of space I shall concentrate henceforth on r^2 rather than ANOVA, but throughout analogous arguments will apply also to the latter, and indeed to other related statistical measures, such as genetic heritability.

In summary, long after Pearson's hostility to causation has been abandoned, still its influence on the *form* of statistical measures of causal efficacy has persisted. The result, I shall argue in this paper, has been baleful.

2. Two Conceptions of Causal Efficacy. How important a cause of heart attacks is a bad diet? Such a question can be understood in two distinct ways: First, how important a cause is diet compared to other causes of heart attacks, i.e., a relative conception; or second, how important a cause is diet in its own right, i.e. an absolute conception. Label these two notions of causal efficacy respectively CE_{rel} and CE_{abs}. Imagine you were wondering whether to improve your diet for the sake of your heart. You might want to know diet's relative importance, and whether it is worth devoting your energies to this rather than to other factors such as exercising or quitting smoking. The CE_{rel} of bad diet would be relevant to this inquiry. Alternatively, you might be interested instead in how much impact fixing diet would have in its own right. For instance, if you have already quit smoking and started exercising, and are pondering now whether to take this extra step. For this latter question, it is a bad diet's CE_{abs} that would be relevant.

One can think of the CE_{rel} sense of causal efficacy as being intimately 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. One can think of CE_{abs}, by contrast, as being intimately associated with *Galilean idealization*, whereby we are concerned with isolating the impact of one factor alone. In this case, far from the causal efficacy being

in part a function of the level of background noise (as with CE_{rel}), we are concerned precisely to discover its value having abstracted all such noise away. How strong, for instance, is the general tendency of a block to accelerate down a slope, independent of 'noise' such as friction and air resistance? Thus CE_{rel} is deliberately a function in part of background noise, whereas CE_{abs} (at least in intention) deliberately is not.

Of our two (as yet somewhat informal) understandings of causal efficacy, the statistical measures r^2 and ANOVA both attempt to capture CE_{rel}. Roughly speaking, for instance, r^2 expresses the proportion of the total (*including* 'background noise') variation of two variables ($\mathrm{Var}(X) \cdot \mathrm{Var}(Y)$) explained just by their covariation ($\mathrm{Cov}(XY)$).[2] In a plot of data points relating two variables X and Y, the r^2 statistic would tell us how tightly those points cluster around a line of best fit. This is appropriate for CE_{rel}. But the CE_{abs} sense of efficacy would be more interested in how much of an increase in Y is yielded by a unit-increase in X. This would at best[3] correspond (for each point) to the slope of the line from that point to the origin. The average CE_{abs} across the sample of data points would then be the average of these slopes, which is independent of how closely those points cluster round a line of best fit. Thus I conclude that r^2 cannot be a good measure of CE_{abs}. This explains why r^2 and ANOVA are commonly used in significance testing,[4] and why they are *not* commonly used in those contexts, such as in physics, where Galilean idealization—and hence use of CE_{abs}—plays a larger role. For similar reasons, it is the CE_{abs} sense (and hence not our statistical measures) that is embraced in much of everyday life too, and also in most of the existing philosophical coverage.[5]

Since the statistical measures therefore capture only one of the two

2. On the relation between variance and effect, see Section 4. ANOVA's formula is (approximately): Var(cause)/Var(total effect), again standardly interpreted as the proportion of total effect explained by the cause.

3. Even this would be neglecting the issue of choice of counterfactual—see Section 6.

4. In this paper I do not address any epistemological issues such as causal inference or hypothesis testing. Rather, the focus is exclusively on the conceptual issue of what we understand by causal efficacy once *given* a cause and effect.

5. Regarding the philosophical literature, see for instance Good 1961 and Miller 1987. The machinery of Bayes nets and causal graphs has also from the start assumed the CE_{abs} understanding of causal efficacy (Spirtes et al. 2000, Pearl 2000). The critiques there of statistical practice mainly concern techniques of causal inference from statistical data, not conceptions of causal efficacy. To my knowledge, Sober et al. 1992 and especially Sober 1988 are the only philosophical treatments that discuss this paper's distinction between CE_{abs} and CE_{rel}.

senses of causal efficacy, I propose an alternative definition to fill the gap.[6] Let the efficacy of a cause C with respect to an effect E be

$$E(C \ \& \ W) - E(C_0 \ \& \ W),$$

where W is background conditions, and C_0 a baseline counterfactual level of C. In the simplest case, C_0 will just be the absence of C. Any value for the efficacy of C will be relativized 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 relativizations is just the logic of controlled experiment—to assess the impact of introducing the factor C, we want to keep constant everything else causally relevant. The second relativization captures the way in which the efficacy ascribed to a cause also depends on what contrast class we are comparing it to. Several further technicalities[7] are inessential here, so I shall gloss over them.

This formula, I claim, captures the CE_{abs} sense of causal efficacy, since consideration of how extraneous factors vary is now explicitly excluded. Notice also how, in direct contrast to r^2 earlier, the variables C and E are treated asymmetrically, reflecting the asymmetry of cause and effect. Explicit reference is made to causation, and the controlled-experiment sensibility with regard to W is unmotivated without it.[8]

3. Are the Two Conceptions Being Confused? So we have two different conceptions of causal efficacy, namely CE_{rel} and CE_{abs}. Standard methodology in biology and psychology, among other sciences, endorses the statistical measures and hence implicitly the CE_{rel} conception. But in these cases, is the CE_{rel} conception always really the one we are actually interested in? Many times, I shall argue, it is not. Accordingly, many times there will be an unstated conflation of the two notions, *implicit* in the presentation of the statistical (and hence CE_{rel}) result when the explanandum in question suggests that we are actually interested instead in the CE_{abs} one. Only a full survey of many applications could demonstrate this conflation conclusively. To make it seem plausible in the space avail-

6. The following formula assumes that cause and effect have already been specified; it is not intended as a definition of causation itself.

7. For instance, strictly speaking the W in the left-hand term is different from that in the right-hand term, since the switch from C_0 to C will in general alter additional things in the world besides our effect of interest. Other omitted technicalities include how to interpret C_0 in cases where the absence of a cause is not well defined, extension to probabilistic rather than deterministic causation, and plenty more besides.

8. Note also that the formula is quantitative, thus allowing *degrees* of causal efficacy. This disarms one of Pearson's major complaints against causation, namely his conception of it as being an unsatisfactorily all-or-nothing affair.

TABLE 1. TOPICS OF r^2 ANALYSES.

First Variable (Presumed Cause)	Second Variable (Presumed Effect)
Percentage of births to unmarried mothers	Percentage of newborns under 2.5 kg
Consumption of saturated fats	Chance of suffering atherosclerosis
Number of voluntary homework problems completed	Score in final exam
Early life exposure to radioactive iodine fall-out from nuclear tests	Later score in SATs
Quantity of smoking	Life expectancy
Quantity of body fat in female professional golfers	Golf score
IQ of college students	GPA
Hours working in front of a computer screen	Subsequent score in a test of depth-perception

able here, I list the subjects of several actual r^2 analyses. In each case, the choices of variable clearly suggest a particular cause and effect relation, and so the degrees of correlation ought indeed to be interpreted as causal efficacies (see Table 1).[9] In each case, are we interested in CE_{rel} or CE_{abs}? Often, we could plausibly be interested in either. The safest judgment is that there is no univocal answer, and that in different circumstances, either conception of causal efficacy has its place. But it already follows immediately that therefore in many of these studies the *wrong* measure of causal efficacy has been used. All, remember, are using r^2. But that is aimed only at CE_{rel}, so in every case where we are actually interested in CE_{abs}, its use was inappropriate.[10]

In the rest of this paper I shall argue that the situation is in fact even worse than that. For even on its home field, so to speak, r^2 is still unsatisfactory. That is, even in those cases where we are indeed interested in CE_{rel}, still r^2 carries other serious drawbacks. Furthermore, our formula for CE_{abs} from Section 2 can be adapted so as also to capture CE_{rel}, and moreover it avoids those drawbacks attending r^2. It should therefore be declared the preferable measure for *both* types of causal efficacy. Similar remarks would apply to ANOVA. This suggests the final conclusion that, with regard to defining causal efficacy, and notwithstanding their widespread use for this purpose, these statistical measures should be discarded.

9. 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 (Howell 1995, Kiess 1996, Sokal and Rohlf 1995) contain the individual references.

10. Sometimes—not always—a slope coefficient for a line of best fit is also presented. This does do more to address the CE_{abs} sense although, assuming the regression is least squares, still the coefficient does not capture a sample's average CE_{abs} exactly. Note also the points of Section 6.

4. First Problem: Levels versus Variances. There exists a second distinction in how we think about causal efficacy, independent of that between CE_{rel} and CE_{abs}, and that is the distinction between levels and variances. Since its significance in this context seems to have been almost nowhere before discussed (Lewontin (1974) mentions it briefly with respect to ANOVA), I shall devote some space to this issue particularly. Consider the efficacy of (the announcement of) an interest rate cut with respect to stock prices. We might be interested either in the direct impact of the announcement on price levels (CE_{abs}), or in the relative importance of that impact compared to all the other determinants of stock prices (CE_{rel}). In both cases, our focus will have been on the *level* of stock prices. But suppose that we were interested not in their level but rather, like perhaps a hedge fund manager, in their volatility. In this case our focus of interest would not be stock price levels, but rather (some function of) the *variance* of those price levels. That is, depending on our interest, we may be concerned either with the level of some effect variable E, or with the variance of that E.

Now, remember that r^2 is couched in terms of variables' variances and covariances. The real problem turns out to be one of r^2's *inflexibility*—it is not that a focus on variance is necessarily always inappropriate, but rather that it *sometimes* will be. And rather as its definition forced r^2 into a commitment always to CE_{rel} rather than CE_{abs}, so now it also forces it into a commitment always to variances rather than levels. (Again, similar remarks apply to ANOVA.)

To see this, consider a numerical illustration. Imagine two teams of five footballers each kicking a ball in turn. The first team is placed on a mountain pass where the wind is gusting capriciously, while the second is in the sheltered prairie below enjoying a steadier breeze. Assume that each team kicks the ball with identical strength (as it were, each boasts an identical distribution of muscular legs). Assume further that the *average* strength of wind gust is the same in each location; the only thing that does differ between the two samples is the *variance* of those gusts. And assume finally that the ball's total acceleration is caused by the kicks and wind gusts and nothing else, and that these two causes compose additively. (See Tables 2 and 3—the r^2 scores in these tables are for the correlation in that sample between the respective input and the total acceleration.)

For each team, what is the average causal efficacy of their kicks? In the absolute CE_{abs} sense, it is clearly 12 in both cases, since for both teams this is the average extra acceleration imparted to the ball by kicking. What is the average CE_{abs} of the wind? Again, the *average* is the same in each sample, namely 15.

TABLE 2. MOUNTAIN TEAM.

Player	Force of Kick	Strength of Wind	Total Acceleration
One	6	15	21
Two	12	25	37
Three	9	10	19
Four	18	5	23
Five	15	20	35
Mean	12	15	27
Variance	18	50	56
r^2 score	.143	.691	. . .

TABLE 3. PRAIRIE TEAM.

Player	Force of Kick	Strength of Wind	Total Acceleration
Six	6	15	21
Seven	12	17	29
Eight	9	14	23
Nine	18	13	31
Ten	15	16	31
Mean	12	15	27
Variance	18	2	17.6
r^2 score	.891	.018	. . .

What of the CE_{rel} sense of efficacy? I suggest a relation between CE_{abs} and CE_{rel} of

$$CE_{rel} = CE_{abs}/E_{tot},$$

where E_{tot} is the total effect. This captures CE_{rel}'s relativization of a cause's own impact to the total amount of 'noise'. Then each team's CE_{rel} of kicking the ball would be: average CE_{abs}/average E_{tot} = 12/27. Likewise, the average CE_{rel} of the wind would be 15/27, and this too is the same for both teams.

In other words, on *either* understanding of causal efficacy the average scores for kicking and the wind are plausibly the same in each sample. However, the r^2 results tell a very different story. Remember, the r^2 between, say, kicking and the total acceleration is supposed to tell us the CE_{rel} of the former with respect to the latter. We have just calculated this CE_{rel} to be 12/27 in both samples. But the r^2 scores for each sample are, respectively, 0.143 and 0.891. Likewise, whereas the CE_{rel} of the wind is 15/27 for both samples, the relevant r^2 scores are 0.691 and 0.018. What has happened?

The answer, of course, lies in the different *variances* of the wind in the two samples. Since the *average* wind strength is the same each time, so also were the average causal efficacies with respect to the *level* of the ball's acceleration. The difference in variances is irrelevant if we are concen-

trating on levels, but *not* if we are concentrating on variance. In the first sample most of the variance in the ball's acceleration is due to variance in the wind rather than variance in the kicking, while in the second it is the other way around. Thus, with respect to the proportion (CE_{rel}) it contributed to the *variance* of the ball's total acceleration, the kicking was not efficacious in the first sample but very efficacious in the second, and it is this asymmetry that is being picked up by r^2. Similarly, the wind's own asymmetric r^2 scores reflect the much greater proportion of the total variance it accounted for in the first compared to the second sample.

So the adequacy of r^2 seems to depend on what we are interested in. We already saw that it only captures CE_{rel} rather than CE_{abs}. Now it is also apparent that it only captures a focus on variance rather than levels. This is a serious restriction since it seems likely that in practice we are usually more interested in the level not variance of an effect, and in all such instances the r^2 measure of causal efficacy is therefore inappropriate. For example, presumably we are less often concerned about the impact of smoking on the *variance* of life expectancy rather than just on life expectancy itself. Nevertheless, what if we did happen to be interested both in CE_{rel} and in the variance of the ball's acceleration? In these favorable albeit unusual circumstances at least, would r^2 not finally capture just what we want? But I shall now argue that, even here, an adjusted version of our own formula is preferable.

5. Adapting Our Formula. Recall from Section 2 our formula for CE_{abs}: $E(C \,\&\, W) - E(C_0 \,\&\, W)$. It is easily converted into a candidate definition of CE_{rel} by normalizing with respect to the total effect, yielding a CE_{rel} of C with respect to an effect E of[11]

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

What of variances and levels? The key is to remember that our formula is defined in terms of an effect term E, not explicitly in terms of *level* of effect. We are therefore free simply to define E to be a variance if desired. For example, rather than set $E =$ the ball's acceleration, we can set $E =$ the variance of the ball's acceleration.

Combining these two maneuvers, we may now calculate, using our adjusted formula, the CE_{rel} scores for each of kicking and wind in the case where we are interested in the variance of the ball's acceleration—

11. Taking C_0 to be zero input, we were in effect already applying this adjusted formula when calculating the CE_{rel} scores of 12/27 and 15/27 earlier.

TABLE 4. CAUSAL EFFICACY SCORES FOR CE_{rel}, E = VARIANCE.

Sample, Cause	Adjusted Formula	r^2 Score
Mountain, kick	.107	.143
Mountain, wind	.679	.691
Prairie, kick	.886	.891
Prairie, wind	−.023	.018

in other words, the exact case where r^2 is supposedly still appropriate.[12] (See Table 4.)

Our formula, once adjusted for the CE_{rel} and effect-as-variance case, therefore captures almost exactly the pattern of the r^2 scores. But notice a wrinkle: Our score in the second sample for the wind is not just very low, as with r^2, but is actually negative. This reflects the fact that the variance of the ball's acceleration would actually be higher without the addition of the wind at all, in other words that in the second sample Var(kicking) > Var(ball's acceleration) and in effect the wind is acting as a mild stabilizer. That is, introducing the wind actually *lowers* the variance of the ball's acceleration. But r^2 is unable to reflect this explicitly (see Section 6 below for more on why). The point is not that we are necessarily interested in this particular nuance, but rather that we *might* be, and only our formula captures it.

Consider now the CE_{abs} scores for the effect-as-variance case. (See Table 5.) In particular, notice another wrinkle when comparing the pattern of these to that of the r^2 scores, namely that the wind now scores higher in the first sample than does the kicking in the second. Intuitively, this is because in the second sample kicking captures a large proportion of a small total variance, whereas in the first the wind—although in absolute terms varying more—now proportionally captures less because the overall total is larger. Thus the CE_{abs} version of our formula is reflecting the

12. Let X = force of kick, Y = strength of wind, and Z = ball's acceleration, and throughout choose a counterfactual of zero input. Then:

Mountain sample:

1. CE_{rel} of kicking = $[\text{Var}(Z|X\ \&\ Y) - \text{Var}(Z|Y)]/\text{Var}(Z|X\ \&\ Y) = (56 - 50)/56 = 0.107$. The CE_{abs} score of kicking, listed in Table 5, is thus $56 - 50 = 6$. The corresponding r^2 score, of 0.143, is taken from Table 2.
2. CE_{rel} of wind = $[\text{Var}(Z|X\ \&\ Y) - \text{Var}(Z|X)]/\text{Var}(Z|X\ \&\ Y) = (56 - 18)/56 = 0.679$. CE_{abs} of wind = $56 - 18 = 38$.

Prairie sample:

1. CE_{rel} of kicking = $[\text{Var}(Z|X\ \&\ Y) - \text{Var}(Z|Y)]/\text{Var}(Z|X\ \&\ Y) = (17.6 - 2)/17.6 = 0.886$. CE_{abs} of kicking = $17.6 - 2 = 15.6$.
2. CE_{rel} of wind = $[\text{Var}(Z|X\ \&\ Y) - \text{Var}(Z|X)]/\text{Var}(Z|X\ \&\ Y) = (17.6 - 18)/17.6 = -0.023$. CE_{abs} of wind = $17.6 - 18 = -0.4$.

TABLE 5. CAUSAL EFFICACY SCORES FOR CE_{abs}, E = VARIANCE.

Sample, Cause	Adjusted Formula	r^2 Score
Mountain, kick	6	.143
Mountain, wind	38	.691
Prairie, kick	15.6	.891
Prairie, wind	−.4	.018

absolute smallness of kicking's variation in the second sample, and r^2 by contrast its *relative* (to the total variation) largeness. We may or may not be interested in the first of these rather than the second. If not, we could simply revert to our formula's CE_{rel} formulation. Indeed with sufficient ingenuity it would no doubt be possible to gerrymander a version of our formula that reproduced the r^2 results exactly. But why bother? What really matters is to get a measure flexible enough to capture always just that in which we are interested antecedently, and for that our formula is much preferable to the less flexible r^2.

6. Second Problem: Counterfactuals. A second key advantage of our formula is that it facilitates flexibility not just with respect to choice of effect term but also with respect to choice of *counterfactual*. In our calculations above, all the counterfactuals were the simplest one—just a factor's absence. But suppose we wished to assess the impact, say, of the mountain team's kicking compared not to its absence but rather compared to the prairie team's? That is, how much difference would it make switching from one team to the other? Of course, by assumption the two teams are identical with respect to kicking, and so such a swap would not make any difference at all. Our formula reflects this naturally and immediately, even in the case most favorable to r^2, namely the CE_{rel} sense of causal efficacy with respect to the variance of the ball's acceleration. For that circumstance, our formula yields for the CE_{rel} of the mountain *rather than* the prairie team:

[(Var (ball's acceleration) given the mountain team & mountain winds)

− (Var (ball's acceleration) given the prairie team & mountain winds)]

/(Var (ball's acceleration) given the mountain team & mountain winds)

$= (56 - 56)/56 = 0.$

Just this kind of calculation is crucial whenever we are considering *interventions*. In particular, here it licenses the (in this case trivial) recommendation that, were we to *replace* the mountain with the prairie team, it would make no difference to the variance of the ball's acceleration up

there. More widely, such analysis of the impact of interventions is crucial to *policymaking*.[13]

Normally at least one term in our formula will be a counterfactual—what *would* happen to E if we changed one input, keeping the others (W) constant? It is crucial when assessing interventions to get the choice of background conditions in this counterfactual exactly right, and that in turn requires our formula's controlled-experiment sensibility. In this example, that meant in particular that we retained the same background profile of wind gusts all the way through the calculation. A second point is that we must also insert the correct C_0 of interest. For instance, moving some other, stronger-legged, team up into the mountains probably would have made a difference after all. A similar story can be told about any intervention—always we wish to assess the likely consequences of that intervention, comparing it either to doing nothing or to the consequences of some alternative intervention. In each case, this implies evaluation of correctly chosen counterfactuals.

More generally, whenever we start to *apply* knowledge via interventions, considerations of causation become inevitable, and Pearson's acausal dream breaks down (Pearl 2000). Neither r^2 nor ANOVA, in contrast to our formula, incorporates the notion of a counterfactual at all—indeed this is one of the positivist virtues Pearson insisted upon. But the cost is to render them unsuitable for assessing interventions. Perhaps it might be thought that just subtracting one team's r^2 score from the other's would yield us straightforwardly the impact of switching between them. But the problem is that in order to set this up correctly in general it is necessary to incorporate a controlled-experiment sensibility, and it is just this that the acausally conceived statistical measures cannot do. To see why not, imagine now analyzing in terms of r^2 our intervention above of using the mountain team in place of the prairie one. If we follow the strategy of just subtracting the r^2 scores from each other, we find:

$$(r^2 \text{ of mountain team's kicking}) - (r^2 \text{ of prairie team's kicking})$$
$$= 0.143 - 0.891 = -0.748,$$

13. Of course, for practical purposes a policymaker would also likely wish to know about underlying mechanisms—*why* did an intervention lead to the quantity of effect that it did? But such considerations are methodological; they do not impinge on the *conceptual* issue of how to define causal efficacy in the first place.

which is of course different from zero. What has gone wrong?[14] The answer is that we have not controlled for varying background conditions, in particular for varying winds. The right-hand term should be what the counterfactual r^2 of the prairie team's kicking *would* have been given mountain winds, but the only r^2 for the prairie team actually available was that for the steady breeze of the prairie. The greatly reduced wind variance in the prairie means that the same kicking profile is much more relatively important there, which is reflected in the r^2 score. But for the purpose of calibrating the intervention of swapping the teams only (and not the winds), this is irrelevant. The result is that crudely comparing r^2 scores cannot deliver the assessment of causal efficacy we require.

The overall picture is that even in their most favorable case—i.e. CE_{rel} and the variance of effect—the statistical measures again fall short, this time over the issue of choice of counterfactual. The underlying reason is their lack of causal sensibility, and hence lack of sensitivity to the importance of keeping background conditions constant.

7. Conclusion. Statistical reports of causal efficacy, if based on r^2 or ANOVA, should be treated with great caution. They offer no reliable substitute for the hard work of establishing quantitative causal results.

Through an appropriate choice of effect term or counterfactual, our own formula is flexible and deft enough to capture whichever particular causal efficacy we happen to be interested in. In contrast, the two statistical measures are unable to capture CE_{abs}. Moreover, their rigidly acausal nature also leaves them inadequate even in the ostensibly friendlier territory of CE_{rel}. First, in causal contexts, an inflexible fixation on variance rather than levels is frequently inappropriate. And second, the lack of a causally inspired controlled experiment sensibility leaves them unable to analyze potential interventions reliably. At heart, as instruments for measuring causal efficacy, they are still suffering from Pearson's wrong turning of 100 years ago.

REFERENCES

Good, I. J. (1961), "A Causal Calculus" parts I and II, *British Journal for the Philosophy of Science* 11: 305–318 and 12: 43–51.

Howell, David (1995), *Fundamental Statistics for the Behavioral Sciences*. Belmont, CA: Duxbury Press.

14. It might justly be objected that in reality no researcher would ever apply r^2 so blindly as we did here. But the point is that—strictly speaking—this is how r^2 *should* be applied if its pretensions to measure causal efficacy are taken at face value. The fact that common sense would lead us to trim away from that thus supports, rather than diminishes, the point.

Kiess, Harold (1996), *Statistical Concepts for the Behavioral Sciences*. Boston: Allyn and Bacon.
Lewontin, Richard (1974), "The Analysis of Variance and the Analysis of Causes", *American Journal of Human Genetics* 26: 400–411.
Miller, Richard (1987), *Fact and Method*. Princeton, NJ: Princeton University Press.
Pearl, Judea (2000), *Causality*. New York: Cambridge University Press.
Pearson, Karl (1911), *The Grammar of Science*. London: A. and C. Black.
Sober, Elliott (1988), "Apportioning Causal Responsibility", *Journal of Philosophy* 85: 303–318.
Sober, Elliott, Erik Olin Wright, and Andrew Levine (1992), *Reconstructing Marxism*. London and New York: Verso.
Sokal, Robert, and James Rohlf (1995), *Biometry: The Principles and Practice of Statistics in Biological Research*. New York: Freeman.
Spirtes, Peter, Clark Glymour, and Richard Scheines (2000), *Causation, Prediction, and Search*. Cambridge, MA: MIT Press.