The Evolution of Altruism: The Sober/Wilson Model*

Peter Gildenhuys †‡

In what follows, I critique the interpretation that Sober and Wilson offer of their group selection model in *Unto Others*. Sober and Wilson mistakenly claim that their model operates as an example of Simpson's paradox and defend an interpretation of their model according to which groups are operated upon by natural selection. In the place of their interpretation, I offer one that parallels the mathematical calculation of the model's outcome and does not depend on the postulation of a force of group selection or a value for group fitness.

1. Introduction. In *Unto Others: The Evolution and Psychology of Unselfish Behavior* (1998), Elliot Sober and David Sloan Wilson make some considerable headway in explaining the evolution of altruistic behavior through the development of a model that specifies circumstances in which there can be selection for altruism. While this advance is significant, Sober and Wilson's definitions both of a "biological group" and of "altruism" remain inadequate. I will point out these flaws as I go along, but my main critical focus will be on Sober and Wilson's interpretation of the mathematical model they put forward to explain the evolution of altruism. No value is given for "group fitness" in the mathematical representation of their model, which leaves their explanation of the evolution of altruism by group selection questionable. I want to contest Sober and Wilson's interpretation of the mathematical model they present in their text by presenting an alternative causal analysis of what is going on when altruism evolves in the circumstances hypothesized by Sober and Wilson. First, I

*Received December 2001; revised August 2002.

†To contact the author, please write to Peter Gildenhuys, Department of History and Philosophy of Science, 1017 CL, University of Pittsburgh, Pittsburgh, PA 15260; e-mail: peg1@pitt.edu.

‡Thanks to David Hull and two anonymous referees for their helpful comments on earlier drafts of this paper.

Philosophy of Science, 70 (January 2003) pp. 27–48. 0031-8248/2003/7001-0004$10.00

lay out the general model of group selection from *Unto Others,* and then I criticize Sober and Wilson's interpretation of how it works. Next, I present my own version of the situation described by Sober and Wilson. Finally, I will argue that there *is* a case for group selection here, or rather selection operating between individual populations, but there is no case for group selection occurring between the subgroups on which Sober and Wilson focus.

2. The General Model of Group Selection. Sober and Wilson eschew the simple model of altruism according to which altruism is sustained in a population because everyone is an altruist. In such a situation, reciprocity is guaranteed straightforwardly among altruists; no one but altruists benefits from altruism. But such models suffer from a couple of familiar objections. First, they do not explain the evolution of altruism, only its perpetuation. Second, were a selfish alternative allele to arise within a population, it would eliminate the altruistic allele by taking advantage of the altruistic behavior of others without itself paying the costs of altruistic behavior.

The Sober and Wilson model is more complex than the simple model and works this way: A population of organisms, only some of which carry genes that code for altruism, is split into subgroups in which interaction among organisms takes place, sometimes for only a small portion of the development of organisms, sometimes for multiple generations. After this, the organisms congregate into a global population before once more being distributed into new subgroups with a different assortment of members. The cycle continues, with new subgroups formed at every cycle. For altruism to evolve within the population, everything depends on the character of the subgroups formed within it. It is only when the subgroups are uneven in terms of their proportion of altruists and nonaltruists that altruists stand a fighting chance.

The most basic version of the model, the one that Sober and Wilson concentrate upon in their exposition, presents a large population split into two subgroups, the members of which interact in some fashion relevant to the fitness of each individual.[1] When the large group divides, two subgroups that are disproportionate in their constitution are formed: one subgroup contains 80 percent altruists and 20 percent nonaltruists, the other contains 80 percent nonaltruists and 20 percent altruists. The altruists acquire an evolutionary edge because, by and large, they help other altruists while the majority of nonaltruists are segregated into a different subgroup. Of course, the nonaltruists do capitalize on the altruism of their

1. Two subgroups, rather than three or more, are easier to deal with mathematically, but nothing depends on the number of subgroups formed out of the global population.

fellows, but too few of them turn up in the same group as the majority of altruists, so they profit less than do other altruists from the altruism of their conspecifics. Because most of the altruists are in a group together with only a few nonaltruists, *on average* the altruists with their altruistic genes finish ahead by co-operating with one another.

As the number of altruists within the global population grows, it becomes more beneficial for the nonaltruists to freeload, so altruism cannot take over a population. Instead, a polymorphism evolves, kept stable by frequency-dependent selection. The more altruists there are, the more it pays to be selfish. The fewer altruists there are, the more it pays to be an altruist, though it is worth remarking that altruists will gain ground against nonaltruists beginning at a very low population density only when such pioneer altruists somehow manage to end up in the same subgroup so as to take advantage of each other's altruism. Early on, Sober and Wilson offer a hypothetical scenario involving the infamous *D. dendriticum* "brain-worm" parasite that shows how altruism could evolve into a stable polymorphism with selfishness beginning from a single mutant altruistic parasite. Sober and Wilson discuss the formation of biased subgroups more generally later on in their work, something I get to below.

Here are the numbers that Sober and Wilson lay out for the simplest version of their group selection model (1998, 25):

	Group 1	Group 2
n	100	100
p	0.2	0.8
Wa	$10 - 1 + 5(19)/99 = 9.96$	$10 - 1 + 5(79)/99 = 12.99$
Ws	$10 + 5(20)/99 = 11.01$	$10 + 5(80)/99 = 14.04$
n'	1080	1320
p'	0.184	0.787
	Global Population	
N	$100 + 100 = 200$	
P	$[0.2(100) + 0.8(100)]/200 = 0.5$	
N'	$1080 + 1320 = 2400$	
P'	$[0.184(1080) + 0.787(1320)]/2400 = 0.516$	

n = number of organisms in a subgroup
p = proportion of subgroup members that are altruistic
Wa = average fitness of altruists
Ws = average fitness of nonaltruists

n' = number of organisms after interaction within subgroups
p' = proportion of subgroup members that are altruistic after subgroup interaction
N = number of organisms in the global population
P = proportion of the global population that is altruistic
N' = number of organisms in the global population after subgroup interaction
P' = proportion of the global population that is altruistic

I have described Sober and Wilson's model as one in which altruism is sustained by a specific sort of group structure allowing altruistic genes to cause the replication of altruistic genes in other organisms through altruistic behavior. The importance of the possibility of altruistic genes causing their replication in this manner is what explains the requirement that the subgroups vary in their proportion of altruists. Altruists must be grouped together such that more altruists benefit from altruistic deeds than do nonaltruists.

This is not how Sober and Wilson understand their model. They claim that two sorts of forces are at work in the above scenario, the force of organismic (they say "individual") selection and the force of group selection. The force of group selection promotes altruism within the subgroups, while the force of organismic selection promotes selfishness. When considering the evolution of altruism, Sober and Wilson take these forces to act in opposition to one another:

> Between-group selection favors the evolution of altruism; within-group selection favors the evolution of selfishness. These two processes oppose each other. If altruism manages to evolve, this indicates that the group-selection process has been strong enough to overwhelm the force pushing in the opposite direction. (1998, 33)

The "force" of group selection acts in opposition to the "force" of organismic selection in just the same way that Newtonian forces can oppose one another. The analogy made by Sober and Wilson (1998, 33) is with individuals pushing upon a billiard ball in different directions. Altruism promotes the "group fitness" of the subgroups, causing them to grow larger at the expense of the organismic fitness of their members, while selfishness promotes organismic fitness, the reproduction of organisms within the subgroup, at the expense of the fitness of the subgroup. When the forces of group selection for altruism and the force of organismic selection for selfishness cancel out, the stable polymorphism is reached.

Sober and Wilson's understanding of their own model suffers from a serious drawback. Despite their repeated use of "group fitness" to describe what is going on in their model, there is no value in their mathematical

analysis for the term. Instead, the fitness calculations are done entirely in terms of the fitness of different trait groups within each subgroup. These trait groups are the population of altruists in subgroup one, the population of nonaltruists in subgroup one, the group of altruists in subgroup two, and the group of nonaltruists in subgroup two.

To show that "group fitness" is never used in Sober and Wilson's calculations, and that they are made instead using fitness values ascribed to trait groups, let's walk through Sober and Wilson's calculations. The average fitnesses for the four trait groups (group 1 altruists, group 1 nonaltruists, group 2 altruists, group 2 nonaltruists) appear in lines three and four. These are calculated by giving everyone a base fitness value of ten units, subtracting the cost of altruism, if performed, and adding the benefits of altruism received. How much fitness is gained by the members of each subgroup from others' altruistic actions is directly proportional to the constitution of the subgroup. More altruists in a subgroup means more benefit for everyone in the subgroup.

The next number to appear in Sober and Wilson's mathematical formulation is n′, reflecting the new size of the subgroups. This figure is generated (1998, 20–21) for each subgroup by multiplying the average fitnesses of the altruists by their frequency within the population, performing the same operation on the nonaltruists, adding the products together, and multiplying the sum by the number of individuals in each subgroup: $n' = n[pWa + (1\text{-}p\text{-}Ws)]$. Notice how the new size of the subgroup must be calculated by adding the growth of each of its trait groups, the altruists and the nonaltruists, separately. The symbol p′ represents the percentage of altruists within the subgroup after interaction among its members. This value is generated by multiplying the original number of altruists within the subgroup (np) by their average fitness and dividing this figure by the total number of members of the subgroup after interaction.

At this point, Sober and Wilson have values for the average fitness of altruists and nonaltruists within each subgroup, the proportion of altruists in each subgroup, and the size of each subgroup. In the last two lines of their mathematical analysis, they use these values to generate a value for the size of the combined group through simple addition, as well as a value for the proportion of the global population that is altruistic. Nowhere does any value for "group fitness" fit into the mathematical representation. Rather, such a value can at best be abstracted by comparing the growth of the group with a high proportion of altruists (group two) with the one that has a low proportion of altruists (group one). The group with more altruists grows larger and hence is "more fit" (Sober and Wilson 1998, 26).

Sober and Wilson call group selection "the *mechanism* that we have proposed to explain the evolution of altruism" (1998, 31; my italics), and also tell us that "if altruism manages to evolve, this indicates that the

group-selection process has been strong enough to overwhelm the force pushing in the other direction" (1998, 33; my italics). The use of multiple terms to capture the role of group selection in the model, along with the aforementioned force analogy, already indicates some confusion over precisely what this role is meant to be. The difficulties with their interpretation become clearer if the parallel they draw between classic formulations of the theory of natural selection and their own interpretation of their model are analyzed.

In a purported analogy with standard formulations of Darwinian natural selection, Sober and Wilson list the conditions that are necessary to bring about an increase the number of altruists in the global population in their model. Three of the conditions are that there be multiple subgroups, that these that vary in their proportion of altruists, and that they periodically subdivide into interacting subgroups before reassembling into a global population (Sober and Wilson 1998, 26). The first two conditions correspond to the conditions in the standard formulation of natural selection in which there must be multiple individuals that vary in their characteristics, while the last condition is a special feature of their model. In an extension of the analogy, they also say that subgroups with more altruists must be more fit than subgroups with fewer altruists, where fitness is understood as the production of more organisms:

> There must be a direct relationship between the proportion of altruists in the group and the group's output; groups with more altruists must be *more fit* (produce more individual offspring) than groups without altruists. (Sober and Wilson 1998, 26)

As it stands, the formulation of this condition is in need of revision. Quite simply, the groups do not produce offspring, or at least if they did, they would produce *offspring groups* rather than individual organisms. But the latter possibility is explicitly denied by Sober and Wilson: it is of crucial importance to the operation of the model that the members of any one set of subgroups formed by periodic subdivision of the global population recombine after interaction into a global population from which new subgroups are formed with a different assortment of members. The subgroups in the model do not autonomously or independently go on to produce the next set of subgroups. Thus, the analogy with Darwinian selection is misplaced since, according to Darwin, individuals that are more fit go on to produce other individuals that are more fit, while in Sober and Wilson's model, individual subgroups that are more fit do not go on to produce individual subgroups that are more fit.

Sober and Wilson cite a final condition for the successful operation of their model:

> To be sufficient, the differential fitness of groups (the force favoring the altruists) must be strong enough to counter the differential fitness of individuals within the groups (the force favoring the selfish types). (1998, 26)

Here confusion has really set in, for it is "fitness," or rather "differential fitness," both of organisms and of groups, that is acting as a force that affects the resultant distribution of altruists in the global population. If this last condition is understood as an oblique reference to the fact that, all things considered, the members of the group with more altruists must produce more offspring than do the members of the group with fewer altruists, then the condition is not tendentious or even interesting. However, the claim that fitness is a causally relevant variable that determines how many altruists and nonaltruists end up in the global population is certainly wrongheaded, since the fitness of the organisms and the groups is calculated in terms of the number of offspring actually produced, as explained in the prior condition cited above. As it stands, the condition is a mere tautology: only if members of one subgroup actually go on to produce more offspring will they produce more offspring. And it would be of no help for Sober and Wilson to claim that what matters is the expected fitness of the organisms or groups, rather than their actual fitness, since we would be left wondering *why* we should expect the results the model produces. Sober and Wilson have not offered us, in the form of a necessary condition, any explanation of why the model produces the results that it does. The model is abstract, what is expected to happen does happen. Other factors that could affect the actual reproduction of the organisms are ruled out of the picture by the authors' explicit assumption of genetic determinism (Sober and Wilson 1998, 22).

What Sober and Wilson seem to be getting at with their last necessary condition for the evolution of altruism in their model is that the fitness losses accrued from altruists' altruistic behavior must be somehow compensated for by a fitness benefit to those same organisms. Group selection is brought into play by Sober and Wilson to fulfill this role. Altruistic genes benefit the group and group selection favors groups that are more fit, so it is by being in the faster-growing group that altruists get compensated for their altruism. But Sober and Wilson fail to provide any more determinate sense of how the force of group selection operates: What does it do that makes groups grow larger? Nor do Sober and Wilson provide any explanation of how the alternative force of differential "group fitness" is detected except through the output of the model itself. While it is clear from the mathematics that one group does in fact grow larger than the other, and that this group is the one with more altruists in it, no explanation of how this happens is forthcoming. And, given the manner in

which the mathematical calculations are done, there is good reason to suspect that the outcome of the model has nothing to do with group fitness at all. It is worth stressing that in Sober and Wilson's model it is individual organisms, not groups or populations, that reproduce and perform altruistic acts. So what we need is an explanation of how altruists are compensated adequately enough for their altruism that appeals to factors that are causally relevant to the reproduction of individual altruists. Group selection as a force, or a mechanism, or a process affecting or involving groups and group-level traits is not even a good candidate for fulfilling this explanatory role because the trait in question, altruism, is a trait that belongs squarely to organisms rather than groups.

So, dispensing with the notion of "group fitness," which remains conspicuously absent from the mathematics, what becomes of Sober and Wilson's interpretation of their model? Are there two distinct forces acting here, organismic selection and group selection, or group fitness and organismic fitness, one favoring the evolution of altruism and the other promoting selfishness? No. Actually, the gene for altruism and the gene for selfishness have different effects depending on the environments in which the altruistic and selfish organisms find themselves. When surrounded by other altruists, altruism causes the reproduction of altruists by increasing the fitness of other altruists who carry the same gene. Surrounded by nonaltruists, altruism is costly, and causes the spread of selfishness within the population. Each gene has distinct capacities to affect fitness whose operations are dependent on the make-up of the surrounding subgroup.

3. Sober and Wilson's Model Explained. Using their mathematical model, I will offer my own detailed explanation of the effects of altruistic behavior on the fitness of each altruist. Altruistic deeds can both cause a net increase in the fitness of altruists or a net decrease in fitness, depending on the circumstances in which they are performed. It is because biased subgroups are formed in Sober and Wilson's model that conditions are such that altruism is perpetuated in the population.

Sober and Wilson introduce their model with reference to a puzzle for probabilistic theories of causality known as Simpson's paradox. This is the problem of spurious correlation. The classic example of spurious correlation comes from Cartwright (1979). It was thought for a while that being a woman caused one to be rejected from the graduate school at Berkeley. But researchers noted that in every department women were no more likely to be rejected than men, the women simply applied to tougher departments. Being a woman was correlated with a variable that actually brought about the effect in question.

Sober and Wilson offer a simplified version of what happened at Berkeley as an analogy to what is going on in their model. They have us imagine

that a hundred individuals, ten men and ninety women apply to a tough department with a thirty percent acceptance rate, while another hundred individuals, this time ten of them women and ninety of them men apply to an easier department with a sixty percent acceptance rate. Neither department discriminates, so thirty-three women are accepted to the graduate school along with sixty men. Sober and Wilson write:

> A bias exists in the two departments combined, despite the fact that it does not exist in any single department, because the departments contribute unequally to the total number of applicants who are accepted. In just the same way, altruists can increase in frequency in the two groups [in their model] combined, despite the fact that they decrease in frequency within each group, because the groups contribute unequally to the total number of offspring.

Sober and Wilson's analogy is misplaced, and to see why, we need to flesh out the supposed parallel in more detail. Sober and Wilson tell us that the bias in the two groups of applicants combined is owing to the fact that each group contributes unequally to the total number of applicants who are accepted. This explanation fits with their explanation of what's going on in their biological model: altruists can survive and reproduce because the subgroup of altruists grows larger faster than thus contributes more members to the global population. But it is easy to offer a different set of numbers that shows a supposed bias in the total number of applicants accepted to Berkeley despite the fact that each department contributes the *same* number of applicants to the total. Consider the following hypothetical situation in which a hundred people are accepted at Berkeley, fifty-six men and forty-four women. Each department contributes fifty applicants to this total. The English department is tough, with a forty percent acceptance rate, while the physics department, having a sixty percent acceptance rate, is relatively easy to get into. All that is left is to gerrymander the remaining numbers. To make things work out, let's say that there are eighty-three applicants to the physics department and a hundred and twenty-five applicants to the English department. Of the applicants to the physics department, seventy-two are men and eleven are women; of the applicants to the English department, thirty-two are men and ninety-three are women. That makes a hundred and four applicants of each gender. The physics department accepts forty-three men and seven women. The English department accepts thirteen men and thirty-seven women. Each department contributes an equal number of individuals to the total number of acceptances, fifty people each. But the total number of acceptances still looks biased, despite the fact that the same number of women applied as men, a hundred and four each. Fifty-six men, and only forty-four women, are accepted to Berkeley.

Whether or not each department contributes unequally to the total number of acceptances is irrelevant to the operation of Simpson's paradox, as the above hypothetical scenario shows. On the above numbers, the graduate school looks biased, despite the fact that neither department is biased, and also despite the fact that each department contributes an equal number of applicants to the overall total of acceptances. Sober and Wilson have misunderstood how Simpson's paradox operates, and it is worthwhile spending some time to clarify just that.

Simpson's paradox is not really a paradox since there is a clear explanation for what is going on in such scenarios. A purported cause, being a man, does not lead to its purported effect, getting accepted at Berkeley, because of the operation of another cause that works to skew the numbers. In the Berkeley case, the alternate cause in operation is applying to easy departments. It turns out that being a man is correlated with applying to easy departments and being a woman is correlated with applying to tough departments. That is what explains the overall bias in the acceptance of candidates to the graduate school. Sober and Wilson explicitly assume genetic determinism in their model (1998, 22), so the only traits that make a difference to the evolution of altruism in their model are the genes for altruism and selfishness along with the two traits, being among mostly nonaltruists, and being grouped with the majority of altruists, that are acquired when the population divides into subgroups. The distribution of these last two traits is what is responsible for the success of the altruists over the nonaltruists. Being surrounded by altruists increases fitness, and altruism is correlated with being surrounded by altruists while selfishness is correlated with being surrounded by nonaltruists.

The upshot of the Berkeley case and other situations in which Simpson's paradox is at work is that a purported cause, a characteristic that looks to be causally connected to another characteristic based on statistical evidence, turns out not to be a cause of the effect at all. Instead, it turns out to be a characteristic that is correlated with another characteristic that is the real cause of the effect.[2] This is not the moral that Sober and Wilson want to draw with respect to their model. The upshot of Sober and Wilson's model is not that altruism does not cause the reproduction of altruists or the replication of altruistic genes. Sober and Wilson's model is like the Berkeley case insofar as there is another cause other than altruism relevant to the effect in question that is at work in their model, namely

2. In order to avoid Simpson's paradox, advocates of probabilistic theories of causality, such as Cartwright (1979) whom Sober and Wilson cite as the source of the Berkeley example, adopt the contextual unanimity clause, stating that causal relations can be inferred from correlations only for those situations that are homogenous with respect to other causes of the effect in question.

the characteristic of being grouped mainly with nonaltruists in subgroup one versus being grouped mainly with altruists in subgroup two. It is also similar insofar as being in group two, a trait that is beneficial to fitness, is correlated with being an altruist, and it is this correlation that explains why altruists outcompete nonaltruists. But Sober and Wilson's model is different from the situation at Berkeley, since it turned out that being a man had no effect at all on the probability of one's being accepted, while altruism does have an effect on the probability of reproduction of altruists. Altruism and selfishness remain variables causally relevant to the reproduction of the organisms that bear the genes that code for these activities. The trick, then, is to handle each of the four populations that are distinct in terms of the variables that are causally relevant to fitness separately, and Sober and Wilson do just that by using the average fitnesses of the altruists and the nonaltruists in each subgroup (*Wa* and *Ws*) separately in their calculations.

Sober and Wilson's model has more to do with another puzzle, Hesslow's (1976) purported counter-example to probabilistic theories of causality, subsequently discussed in a slightly modified form by Cartwright (1989) and Eells (1991). Figure One represents the standard causal path-analysis for the notorious pills/thrombosis case and Figure Two represents what I take to be going on in Sober and Wilson's model:

In both cases, two causal paths, one whose net effect is positive and the

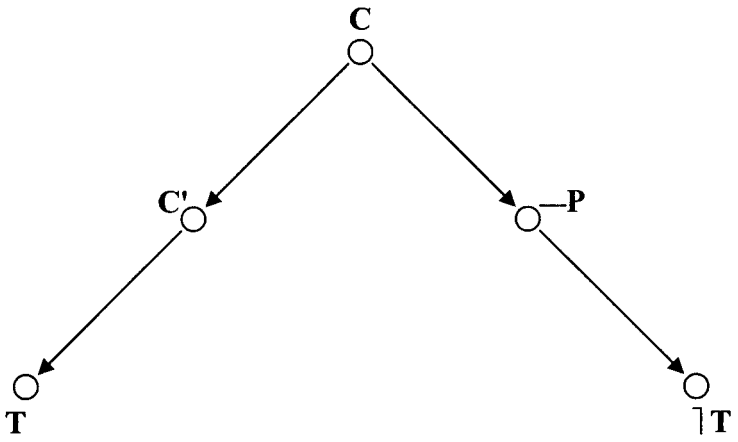

Figure 1. Here is the diagram used by Cartwright (1989, 99) to illustrate the paradoxical effects of taking birth control pills (**C**) on thrombosis (**T**). The path of arrows on the left illustrates the positive causal connection between the pills and thrombosis. The pills alter their takers' blood chemistry, here illustrated as the production of a fictional chemical (**C′**), and this causes thrombosis. The pills also prevent pregnancy (**−P**) which leads to a decrease in the likelihood of thrombosis (**⌉T**), as illustrated by the causal chain on the right.

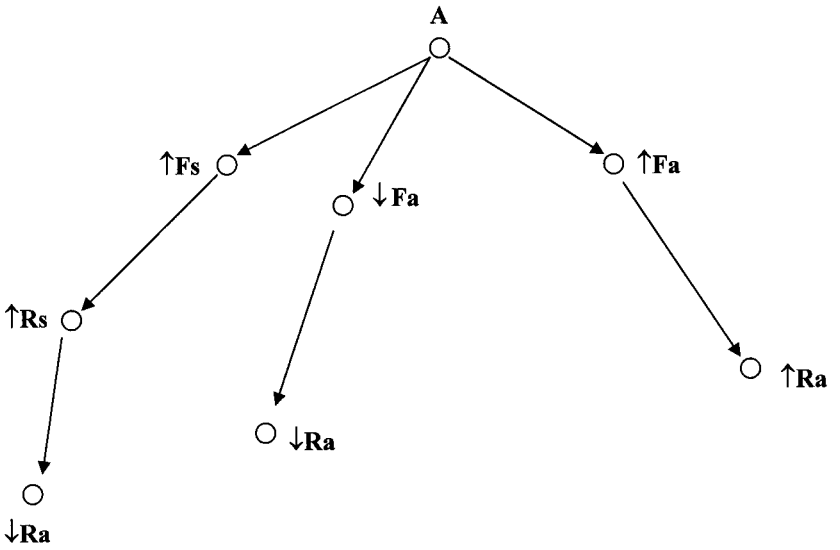

Figure 2. Constructed to resemble Cartwright's illustration of the effects of birth control pills on pregnancy, this diagram shows the effects of altruism (**A**) on the reproduction of nonaltruists and altruists. As illustrated by the left-hand causal chain, altruism increases the fitness of nonaltruists (**↑Fs**) when they benefit from altruistic deeds, and thus increases the likelihood of their reproduction (**↑Rs**). Because the nonaltruists and the altruists are in competition, anything that causes an increase in the likelihood of reproduction of the nonaltruists causes a decrease in likelihood of reproduction to the altruists (**↓Ra**), as shown by the last link in the chain. The middle causal chain shows how altruism decreases the likelihood of reproduction of the altruists in a second fashion: altruistic deeds decrease the fitness of the agents that perform them(**↓Fa**), thereby decreasing their likelihood of reproduction (**↓Ra**). Altruism, when it benefits other altruists, increases their fitness (**↑Fa**) leading to an increase in their likelihood of reproduction (**↑Ra**), as illustrated by the right-hand causal chain. Altruism acts both to bring about and inhibit the reproduction of altruists simultaneously, just as birth control pills simultaneously bring about and inhibit thrombosis.

other negative, link the causes (taking the pills or behaving altruistically) to their effects upon fitness or the probability of acquiring thrombosis. Whether or not birth control pills increase the likelihood of thrombosis or decrease it is obviously dependent on a number of factors, most obviously the pill taker's chance of getting pregnant if she does not take the pills. For instance, if an individual is incapable of getting pregnant, then in her case the preventative causal chain linking the pills to thrombosis is made inoperative and the pills will cause thrombosis. Similarly, the causal effects of an altruistic gene on making copies of itself are dependent on the circumstances in which altruistic deeds are committed. Altruistic genes cause the replication of other altruistic genes when they affect other altruists and fail to do so when the altruism they induce benefits nonaltruists. Indeed, Sober and Wilson's model is not even identical to the

pills/thrombosis case because altruism has three causal consequences, not just two. The altruism of the members of each subgroup produces a fitness increase for altruists when altruists benefit, a fitness loss for altruists when nonaltruists benefit (or conversely, a fitness benefit to nonaltruists who freeload), and a fitness loss to the altruistic agent no matter who benefits from the altruism. Thus, the positive causal path linking altruism to the reproduction of altruists must predominate over both negative causal chains. Selection for altruism can occur only if altruists are grouped together and receive the bulk of each other's altruism. The variable of being grouped with most of the altruists, a characteristic that causes reproduction, must be correlated with being an altruist, a characteristic that inhibits reproduction.

4. How the Model Works. I want to offer another representation of the causal effects of altruism, one that breaks it down into its parts in order to add up the effects of altruism on the fitness of altruists and the effects of altruism on the fitness of nonaltruists. Because altruists are grouped mainly with other altruists in group two, the bulk of the fitness benefits produced by altruists is directed toward other altruists. Even when we factor in the cost of altruism performed, altruists still finish ahead. The net effect of altruism in the Sober and Wilson model is to promote the relative fitness of altruists. Figure Three and Figure Four show the causal effects of altruism in each subgroup in Sober and Wilson's model along with a mathematical representation of the quantity of fitness benefits and costs incurred by altruists and nonaltruists.[3]

In Sober and Wilson's model, altruistic genes indirectly cause the replication of altruistic genes because by and large altruists are grouped with other altruists. Altruism, which decreases the fitness of the agent, is correlated with being in the same group as the majority of altruists, a trait that increases fitness by allowing altruists to take advantage of their fellows' altruism. In the conditions stipulated in Sober and Wilson's model, altruism is ultimately beneficial to the altruists, as is shown by Figure Five representing the effects of altruism in both groups combined.

3. These numbers look different from Sober and Wilson's because I am considering only the causal effects of altruism in each of the two groups, that is, what difference altruism makes to the fitnesses of altruists and nonaltruists. To get back to Sober and Wilson's calculations simply average out the net change to the fitness of the nonaltruists and altruists in each subgroup (net change/number of altruist subgroup members) and add it to ten, the base fitness that Sober and Wilson give to each individual. This will yield Sober and Wilson's values Wa and Ws, the average fitness value for altruists and nonaltruists in each subgroup. Wa in group one: $10 + -1/20 = 9.95$. Wa in group two: $10 + (239)/80 = 12.99$. Ws in group one: $10 + 81/80 = 11.01$. Ws in group two: $10 + 81/20 = 14.05$. Some of the numbers are a hair off because I have rounded off my calculations at different points.

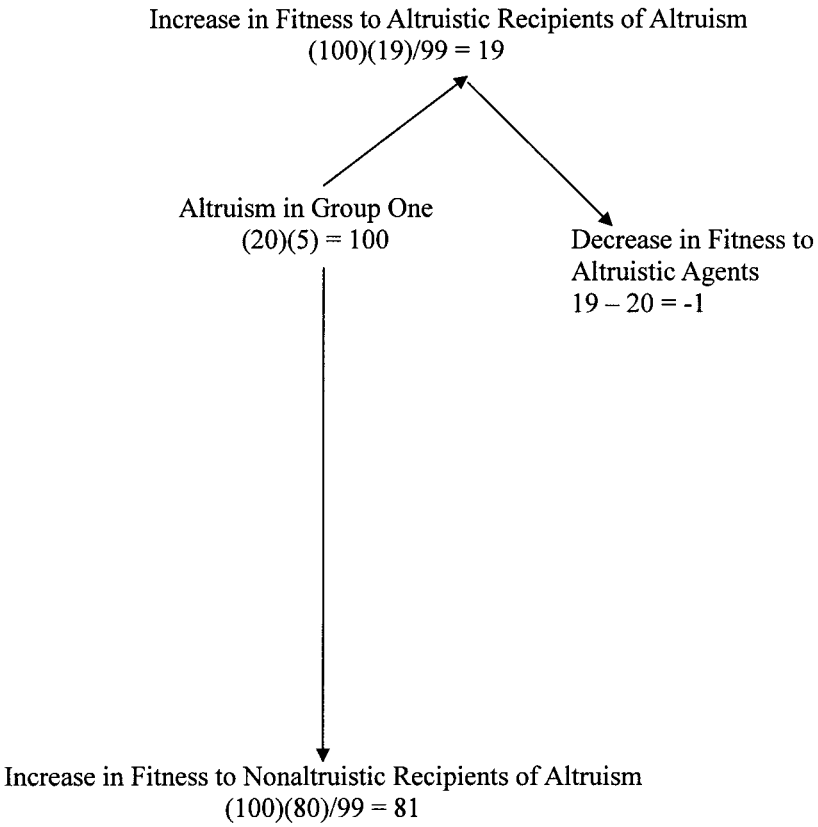

Figure 3. This figure shows the causal effects of altruism in group one on the fitness of altruists and nonaltruists separately. The pool of a hundred units of fitness produced by the altruistic behavior of the twenty altruists in the group is divided up evenly among the other group members. The altruists receive just less than a fifth of the benefits of the altruism (since they cannot benefit from their own altruism) and also suffer a penalty of a single fitness point each for their altruistic sacrifice. In this group, altruism leads to the reproduction of nonaltruists who reap the lion's share of the benefits from the altruistic deeds.

5. Altruism Defined. Once we recognize that altruistic genes must replicate themselves indirectly by benefiting other altruists, Sober and Wilson's definition of altruism comes into question. According to Sober and Wilson altruistic behavior, "increases the [relative] fitness of others and decreases the [relative] fitness of the agent" (1998, 17). I twice added in "relative" because of what Sober and Wilson say later on (1998, 33; italics in original): "In general, evolutionary success depends on *relative* fitness (Williams 1966). It doesn't matter how many offspring you have; it only matters that you have more than anyone else." According to this definition, any gene that causes fitness-reducing behavior will count as altruistic, de-

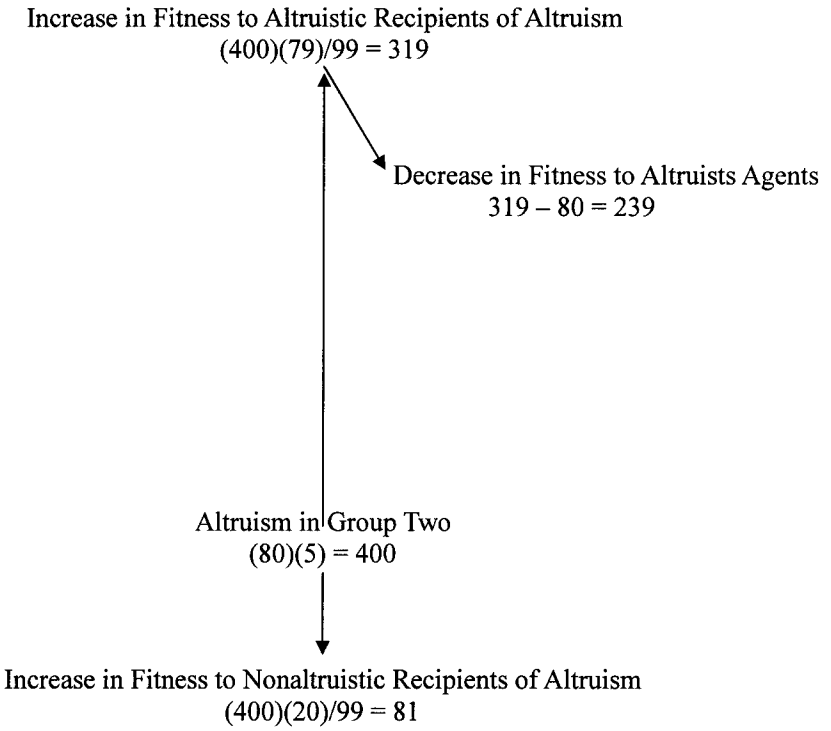

Figure 4. This figure shows the causal effects of altruism on the fitness of altruists and nonaltruists of group two separately. In this group, where altruists predominate, they receive the bulk of the altruism produced by the group's eighty altruists. Even when the fitness loss brought about by their altruistic behavior is factored in, the altruists still fare far better than their nonaltruistic counterparts. Not only is this a group in which the altruists receive the bulk of the fitness benefits from their altruism, this group also produces four times as much fitness benefit to be divided up among the members.

spite the fact that genes that do no more than cause a loss in fitness simply cannot last in the gene pool except by luck. Biting off one's leg is an altruistic behavior, according to Sober and Wilson's definition, because the behavior lowers the relative fitness of the organism that does it and thus increases the relative fitness of other conspecific competitors. But to construe this as biological altruism is a mistake, because such behavior could not evolve even in the situation postulated by Sober and Wilson's model, since nothing is to be gained when a group of organisms all get together and mutilate themselves. Contrary to Sober and Wilson's definition, to count as altruistic an action must do more than simply alter the relative fitness of two agents. It must also be the sort of behavior whose net benefit to other altruists must be greater than its cost to the agent. This condition is satisfied by Sober and Wilson's model because altruistic

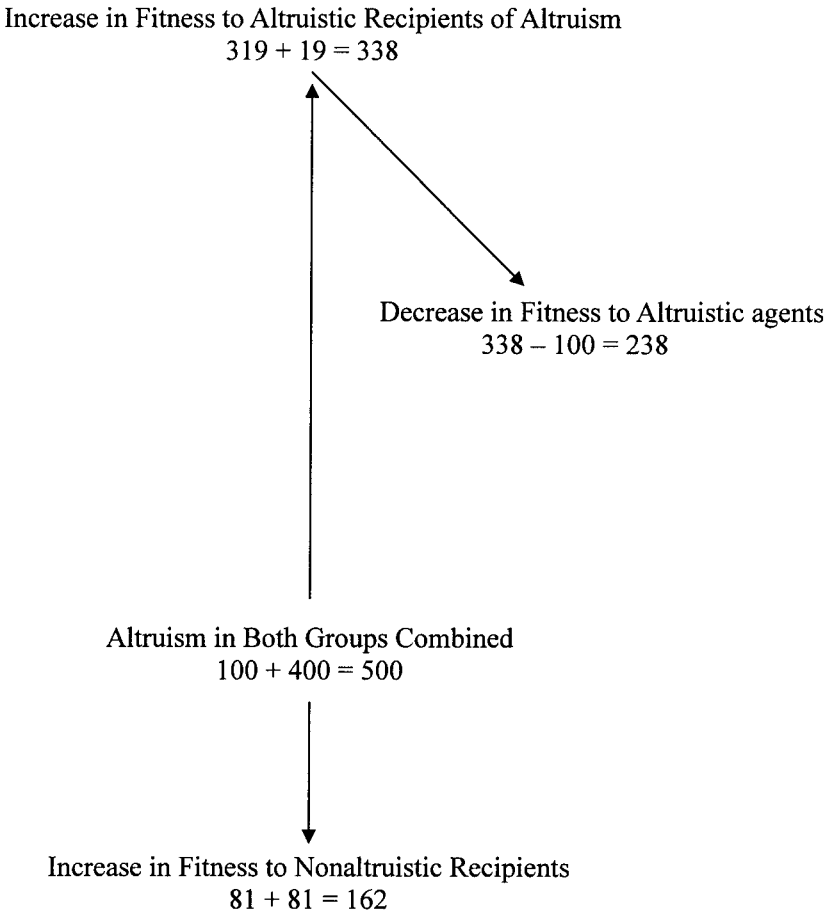

Figure 5. This figure shows the effects of altruism in groups one and two combined. Overall, the altruists gain two hundred and thirty-eight units of fitness from each other's altruism while the nonaltruists gain one hundred and sixty-two fitness units by freeloading (these values are just summed from those in Figures 3 and 4). All other things being equal, the global population, which began with the same number of altruists as nonaltruists, will have become disproportionately altruistic before it subdivides once again. Thus, in the scenario hypothesized by Sober and Wilson, altruism causes the reproduction of altruists by increasing their fitness and hence their probability of reproduction.

genes cause a five point increase in fitness to others, and it is mostly altruists that receive the fitness boost, while costing only a single fitness point deficit to the agent (1998, 25). For altruism to maintain a foothold within a population of organisms, it is enough that on average an altruist is just as likely to gain, in terms of relative fitness to altruists and nonaltruists, from the altruism of other altruists as it is to suffer from its own altruistic actions.

6. Altruism and Averaging. By paying attention to the causal chains linking altruistic behavior with fitness values we can see why Sober and Wilson do their calculations using fitness values ascribed to the trait groups (the altruists and the nonaltruists in each subgroup) in place of the subgroups in their model. Members of the trait groups are all affected in the same manner by the altruistic actions of their counterparts, and they all affect the other members of their subgroup in an identical fashion. In short, the members of each trait group are all exposed to the same set of causally relevant factors and can be treated as a group when calculating fitness values. But this is not the case with the subgroups in Sober and Wilson's model. Here, the groups represent members of the population that interact in ways relevant to organismic fitness, not members of the population that share a common set of factors causally relevant to their reproduction. It is the absence of an overlap between these two sorts of groups that indicates that no "force" of group selection operating between the subgroups is responsible for the evolution of altruism in Sober and Wilson's model. The subgroups are not groups with which one can do causal reasoning; they cannot be treated as a homogenous unit since their members vary in terms of those characteristics that are causally relevant to the probability of their reproduction.

I have focused on the details of the causal chains that link altruistic genes to the perpetuation of altruism because I do not want to be accused of adopting what Sober and Wilson dub "the averaging fallacy" (1998, 31). Sober and Wilson point out that if we simply look at the output of their model, it looks as though altruistic genes are "really" selfish, since they manage to outcompete the nonaltruistic ones (1998, 31). But careful attention to exactly how altruistic genes accomplish their replication shows that they are truly altruistic, at least insofar as this means that they must bring about the reproduction of another altruistic organism in order to cause their replication. Sober and Wilson also write that "another reason to reject the averaging approach is that it fails to identify the separate causal processes that contribute to the evolutionary outcome" (1998, 32). While I disagree with their gloss on what these processes are, I have discussed in detail my alternative representation of their model. Only one feature of the model has been left out of my analysis. The positive causal chain that links altruistic action with increased fitness for altruists dominates over the negative causal chain because the subgroups are disproportionate in terms of the number of altruists in each. What we want to know, then, is how the subgroups become biased in that fashion.

7. How Subgroups Are Formed. Sober and Wilson write: "A group is defined as a set of individuals that influence each other's fitness with respect to a certain trait but not the fitness of those outside the group" (1998, 92). This definition is inadequate for two reasons. First of all, it encompasses

far too much. The biological world is way too complicated for such loosely defined groups to have any discernible role. Just imagine the group of aquatic organisms that influence one another with respect to swimming ability. But this is not the main difficulty with their definition. The definition is misleading because it makes it seem as though the subgroups in the model consist of those organisms that *actually do* interact with one another. This leaves out exactly *why* the organisms interact with one another. The point is that the groups are not composed simply out of interacting organisms, but rather composed out of organisms that, for some other reason, *are constrained* such that they *must* interact (in some important fashion) *only* with one another. The bias in the proportion of altruists in each subgroup only matters because interaction must occur within the boundaries of the subgroups.

Sober and Wilson focus on three explanations for the bias in group constitution so essential to their model. Subgroups may be formed randomly, out of kin relations, or out of mutual recognition among altruists. I'll deal with each of these cases in turn, but much here is left to the imagination of the reader. What matters is that groups assort themselves in a biased fashion, not exactly how they do so. I briefly consider, however, one interesting question: To what entity does one ascribe responsibility for this vital characteristic of subgroup formation? There may just be a case for "group selection" here, but it is not to be found where Sober and Wilson are looking for it.

In the case of random group formation, the mechanism by which altruists are grouped with altruists, and hence disproportionately affect other altruists with their altruism, is not, as one might be tempted to suppose, no mechanism at all. Rather, the mechanism is simply a tendency among members of the global population to sort themselves randomly into subgroups. After all, there are lots of ways they could form subgroups and doing so randomly is only one available option. If subgroups are formed randomly, then altruism will be sustained within the population only if the benefits caused by altruists are high relative to their cost, since it is unlikely that, in general, the groups will vary greatly in terms of the proportions of altruistic members. The more egalitarian the constitution of the subgroups, the more altruism must be beneficial to the fitness of the recipient relative to its cost for the agent, for altruism to remain viable within the population. One of Sober and Wilson's examples of a population in which random group formation promotes the evolution of altruism is the population of various strains of the *mixoma* virus developed to control the rabbit population in Australia (1998, 46). The virus, designed to kill, began to reproduce less quickly after infesting rabbits in the wild and so became less virulent a few weeks after its release into the rabbit population. The decrease in virulence occurred because the virus is spread

by mosquitoes that bite only live rabbits. Viral strains that reproduce less quickly and hence do not kill the rabbits they inhabit as quickly will be more likely to make it into another rabbit. Mosquitoes spread the virus from rabbit to rabbit and the mixture of different strains transmitted to any given rabbit (= subgroup), some more and some less virulent (= selfish), is likely to be random.

Subgroups that are formed out of kin provide the most plausible account of how altruism might evolve according to Sober and Wilson's model. Here, altruists manage to disproportionately affect other altruists by interacting with their relatives who, because of shared heredity, are more likely to be altruists than are unrelated individuals. The formation of subgroups of kin, then, can be used to explain how altruism could evolve within a population by natural selection beginning at very low levels within the population. One of the examples here is the *Dicroelium dentriticum* parasite. Some members of the species invade the brains of ants and cause them to hang about on top of grass blades, where they (and their hosts) are eaten by cattle. This act of suicidal altruism benefits their fellow parasites holed up in the ants' digestive system waiting to continue their lifecycle within the ungulate's digestive system (Sober and Wilson 1998, 18). The parasites reproduce asexually before entering the ant so, as in the case of the wasp discussed above, many of the parasites within the ant carry the same genes.

Finally, the subgroups could be formed by reciprocal recognition among altruists. This is often called the "green beard" effect, so named because, according to a popular hypothetical example, the gene for altruism also allows altruists to recognize each other by some obvious marking, the "green beard." But altruists could perhaps more easily recognize each other as altruists simply by attending to whether or not another organism acts altruistically. Sober and Wilson refer to experimental work on guppies (1998, 46) that shows that guppies choose their associates on the basis of their fellows' previous behavior, though it remains unclear to me why all guppies, not just altruistic ones, do not share the same tendency to group themselves with other guppies who are willing to risk their lives checking predators.

8. Group Selection After All? Sober and Wilson's model for the evolution of altruism requires some mechanism for the periodic assortment of the majority of altruists into the same group as one another. As stressed above, this is necessary because altruists must receive the lion's share of the fitness benefits produced by altruistic deeds. Anything that disturbs the operation of the mechanism by which altruists are grouped together will lead to the disappearance of altruism from the group. So the perpetuation of altruism in the conditions hypothesized by Sober and Wilson is

contingent upon a population-level process that causes the periodic formation of interactive subgroups that are unequal in terms of the number of altruists in each. Thus, my rendition of Sober and Wilson's model is not in any sense reductionist, since it must be acknowledged that group-level processes form a necessary condition for the evolution of altruism. However, the fact that a population-level trait must be in place for the model to work does not mean that altruism itself arises as a result of group selection. There may, however, be a case for population-level selection here, but for the group-level property of biased subgroup formation, not for the individual-level property of altruism.

To see this, we must switch the roles played by altruism and biased subgroup formation such that the latter is a cause and the former a mechanism. Viewed in this way, group selection is established between individual populations of organisms, not for altruism, but instead for periodic biased subgroup formation. Selection will favor those populations of organisms that contain altruists and periodically divide into biased subgroups over populations of organisms that do not contain altruists as well as those that do contain altruists but do not divide into biased subgroups. Take for example some population with the characteristic of periodically fragmenting into kin groups. A population that contains a substantial number of altruists that periodically fragments itself in this fashion would prosper at the expense of populations that are also partially composed of altruists, but which do not fragment themselves in this way. Such a population would also be selected over populations that contain no altruists at all.

The key question to ask, then, is whether the tendency of a population to sort itself into biased subgroups is a characteristic of the population as a whole or a characteristic of individual organisms. This is not always an easy question to answer. When subgroups are biased in their proportion of altruists owing to each altruist's ability to recognize other altruists, it seems pretty clear that the bias is owing to a characteristic of individual organisms. When subgroups are biased in their proportion of altruists because of random subgroup formation processes within the population, we have a clear example of a group-level characteristic that is responsible for the bias. When the bias is owing to kin-group formation, the case seems more ambiguous and one would need detailed examples in order to decide. But if it is a population-level property that is responsible for the formation of biased subgroups, then altruism at the organismic level could count as a condition that must be in place in order for the group-level property of biased subgroup formation to perpetuate itself. So perhaps Sober and Wilson's model is a model of population selection after all, but not a model of group selection for altruism. Instead, Sober and Wilson's model can be

construed as a model showing how a population-level property of biased subgroup formation can be selected for when as least some members of the population are altruists.

Altruism itself is not favored by population-level selection; Sober and Wilson's groups, after all, are not properly speaking altruistic. But group structures are at work in the evolution of altruism, since a peculiar group-level property, biased subgroup formation, is a necessary condition for the perpetuation and spread of altruism within the population. And conversely, the peculiar property of biological altruism among members of a population is a necessary condition for the selection of populations that assort themselves into subgroups made up of unequal numbers of altruists, over populations that lack at least one of these characteristics. Sober and Wilson have simply confused the importance of a group-level property for the evolution of altruism with the group selection for altruism.

It is no surprise that Sober and Wilson's model should turn out to be one that shows the possibility of group selection, since it relies on such a peculiar group-level trait, namely the periodic assortment of population members into biased subgroups. In calling their model one that shows the possibility of group selection, I mean to say that they have shown how some irreducibly population-level property, such as periodic random subgroup formation, could be selected over an alternative irreducibly population-level property, such as not forming temporary subgroups at all. Group selection, as I understand it, is a process by which groups with one property outcompete groups with a different property, by increasing the number of groups with the superior trait at the expense of the groups with its inferior alternative. Group selection on my understanding is a matter of the selection of some group-traits over others; what makes it *group* selection rather than *organismic* selection is a matter of which sort of entity has the trait.

9. Conclusion. To sum up, I have argued the following. First, what counts as altruistic behavior, at least of the sort that stands a fighting chance of evolving by natural selection, is not merely behavior that is detrimental to the fitness of the agent. To evolve by natural selection, altruistic behavior must benefit others above and beyond the benefits that they accrue from the corresponding fitness loss on the part of the altruist. Second, when altruism does evolve it evolves because, on average, altruistic genes indirectly benefit other altruistic genes more than they benefit nonaltruistic genes. Altruistic behaviors are ones that increase fitness indirectly, through benefiting other individuals within a population who share an altruistic disposition. Third, group selection is not implicated in the evolution of altruism on Sober and Wilson's model. But, though they did not realize

it, Sober and Wilson's model does show how selection could favor some individual populations of organisms that contain altruists and periodically divide into biased subgroups over other populations that contain altruists but do not divide into biased subgroups. Such a process can legitimately be called *group selection.* It is worth remarking, however, that none of the criticisms I have made of Sober and Wilson's model is aimed at refuting it. The model stands as a powerful instrument for explaining the evolution of altruism. Instead of refuting the model, I have shown that group selection is not implicated in its workings in the manner that Sober and Wilson suggest.

REFERENCES

Cartwright, Nancy (1979), "Causal Laws and Effective Strategies", *Nous* 13: 419–437.
——— (1989), *Nature's Capacities and Their Measurement.* Oxford: Clarendon Press.
Eells, Ellery (1991), *Probabilistic Causality.* Cambridge: Cambridge University Press.
Hesslow, Germund (1976), "Discussion: Two Notes on the Probabilistic Approach to Causality", *Philosophy of Science* 43: 290–292.
Sober, Elliott, and David Sloan Wilson (1998), *Unto Others: The Evolution and Psychology of Unselfish Behavior.* Cambridge: Harvard University Press.
Williams, George (1966), *Adaptation and Natural Selection: A Critique of Some Current Evolutionary Thought.* Princeton: Princeton University Press.