Hamilton's Rule and its Discontents

Jonathan Birch

(forthcoming in *The British Journal for the Philosophy of Science*)

ABSTRACT

In an incendiary 2010 *Nature* article, M. A. Nowak, C. E. Tarnita and E. O. Wilson present a savage critique of the best known and most widely used framework for the study of social evolution, W. D. Hamilton's theory of kin selection. Over a hundred biologists have since rallied to the theory's defence, but Nowak et al. maintain that their arguments 'stand unrefuted'. Here I consider the most contentious claim Nowak et al. defend: that Hamilton's rule, the core explanatory principle of kin selection theory, 'almost never holds'. I first distinguish two versions of Hamilton's rule in contemporary theory: a special version (HRS) that requires restrictive assumptions, and a general version (HRG) that does not. I then show that Nowak et al. are most charitably construed as arguing that HRS almost never holds, while HRG buys its generality at the expense of explanatory power. While their arguments against HRS are fairly uncontroversial, their arguments against HRG are more contentious, yet these have been largely overlooked in the ensuing furore. I consider the arguments for and against the explanatory value of HRG, with a view to assessing what exactly is at stake in the debate. I suggest that the debate hinges on issues concerning the causal interpretability of regression coefficients, and concerning the explanatory function Hamilton's rule is intended to serve.

1 *Nature* Red in Tooth and Claw

In August 2010, in an incendiary *Nature* article entitled 'The Evolution of Eusociality' ([2010]), the Harvard sociobiologists Martin A. Nowak, Corina E. Tarnita and Edward O. Wilson unleashed a savage critique of the best known and most widely used framework for the study of social evolution, W. D. Hamilton's ([1964], [1970]) theory of kin selection. The article sparked a ferocious controversy. In March 2011, *Nature* published five rebuttals[1], one of which remarked in no uncertain terms that Nowak and colleagues' arguments 'are based on a misunderstanding of evolutionary theory, and a misrepresentation of the empirical literature' (Abbot et al. [2011], p. E1). The letter was signed by 137 social evolution theorists. Further rebuttals have followed[2], but Nowak, Tarnita and Wilson remain unmoved. In an online statement dated June 2011, they write:

> Some of the criticism distorts our arguments, which should remain clear. We therefore provide a brief summary of our main points, all of which stand unrefuted... (Nowak et al. [2011b])

In this paper, I want to deconstruct this rather acrimonious debate. I want to suggest why it has reached the present state of deadlock, and how the deadlock might nevertheless be broken.

[1] (Abbot et al. [2011]; Boomsma et al. [2011]; Strassmann et al. [2011]; Ferriere and Michod [2011]; Herre and Wcislo [2011]); see (Nowak et al. [2011a]) for the authors' uncompromising reply.

[2] See, e.g., (Rousset and Lion [2011]; Gardner et al. [2011]; Bourke [2011a]). Not all responses, however, have been negative: see, e.g., (Doebeli [2010]; van Veelen et al. [2010]).

I will focus on what I take to be the core disagreement between Nowak, Tarnita and Wilson and their opponents. The disagreement concerns *Hamilton's rule*.[3] The rule states, broadly speaking, that a social behaviour will be favoured by natural selection if and only if $rb - c > 0$, where *b* represents the 'benefit' the behaviour confers on the recipient, *c* represents the 'cost' it imposes on the actor, and *r* represents the 'relatedness' between actors and recipients. Note that, although talk of 'costs' and 'benefits' intuitively connotes that costs will detract from an agent's fitness while benefits increase it, this need not be the case: the rule is intended to apply regardless of the sign of *b* or *c*. Hence, while the rule is most often associated with the evolution of cooperation (for which *b* is positive) and the evolution of altruism (for which *b* and *c* are *both* positive), selfish, spiteful and mutualistic behaviours are also intended to fall within the scope of Hamilton's rule (see Hamilton [1964]; Trivers [1985]; Bourke and Franks [1995]; West et al. [2007]; Bourke [2011b]). Note also that *r* need not measure *genealogical* relatedness. Though Hamilton's rule is often glossed informally in terms of genealogical kinship, it has long been recognized that the rule can still apply when genotypic or phenotypic correlations arise by other means (see, e.g., Hamilton [1975]; Grafen [1985]).

Nowak, Tarnita and Wilson provocatively assert that Hamilton's rule 'almost never holds' (Nowak, Tarnita and Wilson [2010], p. 1059), in the sense that it almost never constitutes a true statement of the conditions under which a social behaviour will be favoured by natural selection. More than any other in the paper, this claim has elicited vigorous

[3] The debate has other facets that I do not discuss here. For instance, I do not discuss the notion of 'inclusive fitness', its relationship to Hamilton's rule, or the 'organism as inclusive-fitness-maximizing agent' analogy it is often thought to underwrite (Grafen [2006]). These are important issues, but I leave them for another occasion.

rebuttals from their opponents—most notably from the Oxford theorists Andy Gardner, Stuart A. West and Geoff Wild, who retort that 'it is simply incorrect to claim that Hamilton's rule requires restrictive assumptions or that it almost never holds' (Gardner et al. [2011], p. 1038). There is, at present, no sign of an end to this standoff. Harvard's most eminent social evolution theorists say that Hamilton's rule 'almost never holds'; Oxford's most eminent social evolution theorists dismiss this claim as 'simply incorrect'. It is hard to see how they can both be right, yet neither seems likely to budge. Is there any way to bring the two opposing camps together?

Here is a conciliatory suggestion. I think that what this debate has brought to the surface is that there are at least two versions of Hamilton's rule in contemporary social evolution theory. Though similar on the surface, these two versions differ significantly in their underlying features; and which version of the rule one has in mind will affect one's views regarding how generally the rule holds. Failure to attend to the different versions of Hamilton's rule has resulted in the Harvard and Oxford camps talking past one another. To make any progress in this debate, we must take greater care to distinguish them.

.The outline of the paper is as follows. In Section 2, I characterize the two versions of Hamilton's rule at issue in the current debate. The first is a 'special' version in which the 'cost' and 'benefit' terms represent fecundity payoffs in an evolutionary game; the second is a 'general' version, derived from the Price equation, in which the 'cost' and 'benefit' terms are partial regression coefficients. In Section 3, I explain why one's views as to how generally Hamilton's rule holds will depend on which version one has in mind. This leads naturally to the suggestion that Nowak et al. are attacking a straw man—that they show

Hamilton's rule 'almost never holds' only by uncharitably construing it in a particularly narrow sense. In Section 4, however, I argue that this characterization of their argument is misleading: Nowak and colleagues also offer arguments against the general version of Hamilton's rule—arguments which have been largely overlooked by their opponents. I suggest that, in light of this, their overall strategy is best interpreted as that of posing a dilemma for the defender of Hamilton's rule: they argue that the special version of Hamilton's rule lacks wide applicability, while the general version lacks explanatory power. If their arguments work, neither version of Hamilton's rule constitutes the widely applicable explanatory principle its advocates take it to be. In Sections 5 and 6, I consider how defenders of Hamilton's rule can respond to this challenge. The debate, I argue, ultimately turns on philosophical issues concerning the causal interpretability of regression coefficients, and concerning the explanatory function Hamilton's rule is intended to serve.

2 Two Versions of Hamilton's Rule

Disagreement about the uses and limits of Hamilton's rule has been a mainstay of theoretical biology since the 1960s. The reason, in a nutshell, is that Hamilton ([1964]) first derived a result of the form '$rb - c > 0$' in a one-locus population-genetic model that made a number of substantive modelling assumptions, including weak selection and the additivity of fitness effects. In the following decades, theorists (including Hamilton himself) explored the extent to which Hamilton's original assumptions could be relaxed. The upshot was a variety of

routes to '$rb-c>0$'-type results, often with incompatible implications about the conditions under which the result obtains.[4]

Perhaps these disagreements are substantive: perhaps there is only one biologically significant result of the form '$rb-c>0$', and only one true statement of the conditions under which this inequality predicts the evolution of social behaviour. But there is another possibility: it may be that there is more than one biologically significant result of the form '$rb-c>0$', and that the similarity in the surface form of these results masks underlying differences in their biological meaning. If this is correct, the appearance of disagreement over the uses and limits of 'Hamilton's rule' may arise simply from a failure to disambiguate different versions of the rule.

I want to explore this second possibility. For I contend that the method one uses to derive 'Hamilton's rule' will often have a significant impact on the meaning one attaches to its terms; the result is that talk of 'Hamilton's rule' ambiguously refers to any of a number of superficially similar principles. In this section, I illustrate this phenomenon by considering two routes to Hamilton's rule in contemporary evolutionary theory: a route that proceeds via evolutionary game theory, and a route that proceeds via George R. Price's ([1970], [1972]) covariance selection mathematics. I show that, while both methods may be used to derive a result with the surface form of Hamilton's rule, the meaning attached to the terms is very different in the two cases. The rationale for picking out these two derivations, rather than any others in the literature, is that understanding the difference between the versions of

[4] See, e.g, (Hamilton [1970], [1972], [1975]; Orlove [1975]; Charnov [1977]; Charlesworth [1980]; Uyenoyama and Feldman [1980], [1981], [1982]; Uyenoyama et al. [1981]; Michod [1982]; Toro et al. [1982]; Grafen [1985]).

Hamilton's rule they yield is crucial if we want to understand the present Harvard/Oxford standoff.

The key conceptual difference between the two versions may be simply expressed. One construes the B, C and R terms as parameters of an evolutionary game, and represents a condition for the evolution of a social behaviour within a very restricted class of populations. The other construes b, c and r as linear regression coefficients, and represents a highly general condition under which a social behaviour will be favoured by natural selection. Hence, although the two rules have the same surface form, their terms are defined in very different ways; and it is vital that we understand these differences if we want to make sense of the current debate. At bottom, the dispute concerns what we have to sacrifice in return for the generality afforded by the second, more widely applicable version of Hamilton's rule—and whether the price is worth paying.

2.1 The special version (HRS)

The game-theoretic route[5] to a version of Hamilton's rule begins with a simple evolutionary game: the one-shot, two-player Prisoner's Dilemma.[6] We can represent the dilemma with the following payoff matrix, in which B is a benefit conferred on the recipient by the social behaviour under study, and C is a cost incurred by the actor:

[5] The order in which I present the two routes is arbitrary: it is not intended to imply that either version has historical or explanatory priority.

[6] See, e.g., (Queller [1984]; Nowak [2006]; Taylor and Nowak [2007]; Okasha [2008]; van Veelen [2009]; van Veelen et al. [2012]) for variants of this route to Hamilton's rule.

	COOPERATE (Co_2)	**DEFECT** (De_2)
COOPERATE (Co_1)	$B - C$	$-C$
DEFECT (De_1)	B	0

In the Prisoner's Dilemma, the Pareto optimal outcome is mutual cooperation. Unfortunately, in the absence of correlated interactions, DEFECT will always secure a higher expected payoff whenever COOPERATE is costly to perform. To see this, note that the expected payoff for cooperation (W_{Co}) and defection (W_{De}) may be written as functions of B, C and the conditional probability that one's opponent will cooperate:

$$W_{\mathrm{Co}} = \mathrm{P}\left(\mathrm{Co}_2 \mid \mathrm{Co}_1\right)B - C$$
$$W_{\mathrm{De}} = \mathrm{P}\left(\mathrm{Co}_2 \mid \mathrm{De}_1\right)B$$

Uncorrelated interactions implies that $\mathrm{P}\left(\mathrm{Co}_2 \mid \mathrm{Co}_1\right) = \mathrm{P}\left(\mathrm{Co}_2 \mid \mathrm{De}_1\right) = f_{\mathrm{C}}$, where f_{C} is the overall frequency of cooperators in the population. From this assumption it follows that:

$$W_{\mathrm{Co}} > W_{\mathrm{De}} \text{ iff } C < 0$$

Since, by definition, an altruistic behaviour imposes a non-negative cost on the actor, the implication is that altruism cannot evolve in a one-shot two-player Prisoner's Dilemma with uncorrelated interactions. This is one way of conceptualizing the 'problem of cooperation' within the framework of evolutionary game theory (see, e.g., Nowak [2006]; Nowak and Highfield [2011]).

The picture changes when we introduce correlated interactions: that is, when we raise the probability that cooperators will interact with other cooperators, and that defectors will interact with other defectors.[7] We can express this formally by adding an *R*-term to our expressions for the relevant conditional probabilities, where *R* is a parameter that specifies the differential probability that one's opponent will play the same strategy as oneself:

$$\mathrm{P}\left(\mathrm{Co}_2 \mid \mathrm{Co}_1\right) = \left(1 - R\right) f_{\mathrm{Co}} + R$$
$$\mathrm{P}\left(\mathrm{Co}_2 \mid \mathrm{De}_1\right) = \left(1 - R\right) f_{\mathrm{Co}}$$

By plugging these new expressions into the payoff functions that determine the values of W_{Co} and W_{De}, we obtain a new, rather more promising condition for the evolution of cooperation:

$$\textbf{(HRS)}\ W_{\mathrm{Co}} > W_{\mathrm{De}} \text{ iff } RB - C > 0$$

[7] See (Skyrms [1996]) for broader discussion of how evolutionary games are transformed by the introduction of correlated interactions.

In essence, the condition states that altruism *can* evolve in a one-shot, two-player Prisoner's Dilemma, provided the differential probability that an altruist has of interacting with another altruist exceeds the cost-benefit ratio. For reasons that will soon become clear, I will refer to this as the *special version of Hamilton's rule (HRS).*

2.2 The general version (HRG)

The second route to a principle with the form of Hamilton's rule begins with the *Price equation*, a fully general but highly abstract description of evolutionary change:[8]

$$\Delta\bar{g} = \frac{1}{\bar{w}}\left[\mathrm{Cov}\left(w, g\right) + \mathrm{E}\left(w\Delta g\right)\right]$$

The equation states that the overall change in the additive genetic value[9] of a trait, g, from one generation to the next is equal to the sum of two quantities: the normalized covariance between g and fitness (w), and the fitness-weighted expectation of the change in g between a

[8] Hamilton ([1970]) was the first to see the relevance of the Price equation to kin selection theory. The derivation of Hamilton's rule I present here, however, is owed originally to Queller ([1992a]). For similar derivations that proceed via the Price equation, see (Grafen [1985]; Frank [1998]; Gardner et al. [2007]; Gardner [2008]; Gardner and Foster [2008]; Wenseleers et al. [2010]; Gardner et al. [2011]; Marshall [2011a]; Queller [2011]; Damore and Gore [2012]).

[9] An individual's 'additive genetic value' (or 'breeding value') for a particular character is its value for that character as predicted by a linear combination of its alleles, weighted by their average effects on the character (see Queller [1992a], [1992b]; Falconer and Mackay [1996]; Gardner et al. [2007]; Gardner et al. [2011] for further detail). If a behaviour is fully determined by the presence or absence of a single social allele, the additive genetic value with respect to this behaviour is 1 for bearers of the allele and 0 otherwise. In this special case, $\Delta\bar{g}$ is simply the overall change in the frequency of this allele.

parent and its offspring (I will not derive the equation here; see Price [1970], [1972]; Frank [1995], [1998], [2012]; Rice [2004]; Okasha [2006]; Gardner [2008]).

The terms in the Price equation are often given an informal causal gloss: $\mathrm{Cov}(w,g)$ is usually interpreted as the partial change attributable to natural selection, while $\mathrm{E}(w\Delta g)$ is interpreted as the partial change attributable to biased transmission (see Frank [1995], [1998], [2012]; Gardner [2008]; for criticism of this gloss, see Okasha [2006]). The route from the Price equation to a version of Hamilton's rule begins by leaving aside the second term, so as to focus purely on the partial change attributable to natural selection ($\Delta_s \bar{g}$):[10]

$$\Delta_s \bar{g} = \frac{1}{\bar{w}}\left[\mathrm{Cov}(w,g)\right]$$

To derive a version of Hamilton's rule, we partition this covariance term into two components: one corresponding to *rb*, the other to *–c*.[11] The first step is to write a linear regression model for the fitness of an arbitrary individual. This expresses its fitness as a linear function of its own genetic value (g) and the average genetic value of its social partners (g'

[10] See, e.g., (Queller [1992a]; Gardner et al. [2011]). Frank ([1998]) derives a variant of Hamilton's rule that takes into account the effect of biased transmission, but this rule lacks the famous '$rb-c>0$' form.

[11] The general strategy of partitioning the overall covariance between genotype and fitness is also central to multi-level selection theory (Okasha [2006]). There is a particularly strong affinity between the regression route to Hamilton's rule and the 'contextual analysis' approach to multi-level selection (Heisler and Damuth [1987]; Damuth and Heisler [1988]). Both partition the Price equation through regression analysis; the major difference is that, while kin selection includes the genotype of an individual's social partner among the predictors in the regression model, contextual analysis considers properties of the whole group of which the individual is a member.

), in which each quantity is weighted by a partial regression coefficient (for example, $\beta_{w,g.g'}$ represents the partial regression of w on g, correcting for g'):[12]

$$w = \alpha + \beta_{w,g.g'} g + \beta_{w,g'.g} g' + \varepsilon$$

The α-term is a constant, and denotes the intercept of the regression line. The residual, ε, quantifies the extent to which the focal individual's fitness deviates from the regression line. It is important to see that, because the regression equation includes a residual term, it is compatible with *any* set of population data.[13] Naturally, the regression line will fit some populations better than others (i.e., with smaller residuals), but there can be no individual, real or modelled, of which the equation is *false*.

We then substitute the linear regression equation into the Price equation, obtaining the following rather unwieldy partition:

$$\Delta_s \bar{g} = \frac{1}{\bar{w}} \left[\mathrm{Cov}(g,\alpha) + \beta_{w,g.g'} \mathrm{Var}(g) + \beta_{w,g'.g} \mathrm{Cov}(g,g') + \mathrm{Cov}(g,\varepsilon) \right]$$

[12]This can be visualized intuitively as the slope of the line of best fit when one plots w against g while 'holding fixed' g'. Note, however, that we are not literally 'holding fixed' anything: we are correcting for a correlated variable by statistical means, in order to minimize the sum-of-squares of residuals in the regression model as a whole (for further detail, see Lande and Arnold [1983]).

[13] A proviso: the partial regression coefficients are defined only if (i) there is non-zero variance in both predictor variables and (ii) the two predictor variables are not perfectly collinear. These are fairly minimal conditions, and it is reasonable to assume that they will be met in a very wide range of cases.

We can simplify this partition by noting firstly that g cannot co-vary with the intercept of the regression line (i.e., $\mathrm{Cov}(g,\alpha) = 0$), and secondly that it cannot (given standard least-squares theory) co-vary with the residuals in the regression model (i.e., $\mathrm{Cov}(g,\varepsilon) = 0$):[14]

$$\Delta_s \bar{g} = \beta_{w,g.g'} \mathrm{Var}(g) + \beta_{w,g'.g} \mathrm{Cov}(g,g')$$

We then rearrange, and exploit the fact that $\mathrm{Cov}(g,g') / \mathrm{Var}(g) = \beta_{g'.g}$, to obtain the following expression:

$$\Delta_s \bar{g} = \frac{1}{\bar{w}} \left[\left(\beta_{w,g.g'} + \beta_{w,g'.g} \cdot \beta_{g',g} \right) \cdot \mathrm{Var}(g) \right]$$

Finally, we exploit the fact that neither $\bar{w}$ nor $\mathrm{Var}(g)$ can be negative to obtain the following rule:

$$\Delta_s \bar{g} > 0 \text{ iff } \beta_{w,g.g'} + \beta_{w,g'.g} \cdot \beta_{g',g} > 0$$

This result already has the form of Hamilton's rule. By re-labelling the regression coefficients as $-c$, b and r respectively, and by swapping the order in which the terms appear, we obtain the rule in a more familiar guise:[15]

[14] Note that the covariance between g and itself is simply the variance in g (i.e., $\mathrm{Cov}(g,g) = \mathrm{Var}(g)$).

[15] This is the 'direct fitness' route to Hamilton's rule; a very similar result may be obtained by considering the effects of a behaviour on the actor's 'inclusive fitness' (see Frank [1998]; Gardner et al. [2011]; Queller [2011]). I use the 'direct fitness' derivation because the 'inclusive fitness' derivation adds some complications which are unnecessary for the purposes of this paper.

$$\textbf{(HRG)} \quad \Delta_s \bar{g} > 0 \text{ iff } rb - c > 0$$

I will refer to this as the *general version of Hamilton's rule (HRG).* As in the case of HRS, the reasons for this label will soon become clear.

2.3 The rules compared

Let us review. The special version of Hamilton's rule, HRS, is a game-theoretic result derived in the context of a two-player, one-shot Prisoner's Dilemma with correlated interactions. The *C* and *B* terms are fecundity payoffs, while *R* is a further parameter that determines the differential probability that social partners play the same strategy. By contrast, HRG has no essential ties to evolutionary game theory. In HRG, the *b* and *c* terms are partial regression coefficients that quantify the overall statistical dependence of one's fitness on, respectively, one's own genotype and that of one's social partners; while *r* is the simple regression of one's social partners' average genetic value on one's own genetic value.[16] In deriving HRG, we made no substantial assumptions about the population or model under study. All we needed were two equations—the Price equation and a linear regression

[16] If each agent's strategy is fully determined by its additive genetic value, then, in the one-shot, two-player Prisoner's Dilemma, the regression of one's partner's genetic value on one's own genetic value is equal to the *R*-parameter, i.e., $r = R$. Importantly, however, the two notions are *conceptually* distinct despite their numerical equality in this particular case. *R* is a model parameter that determines an aspect of population structure, whereas *r* is a population statistic that measures the overall association between the genotypes of social partners.

equation—which are true of virtually any evolving population, whether real or modelled, and from which HRG follows a priori.

3 How They Come Apart: A Simple Illustration

The differences between HRS and HRG are not merely superficial: they have significant implications regarding the conditions under which the two rules hold. This makes the distinction crucial for understanding the current debate. For when Nowak, Tarnita and Wilson assert that Hamilton's rule 'almost never holds' ([2010], p. 1059), it is clear from the way in which they formalize their arguments that HRS, not HRG, is the version of Hamilton's rule they have in mind. Meanwhile, when Gardner, West and Wild reply that 'it is simply incorrect to claim that Hamilton's rule requires restrictive assumptions or that it almost never holds' ([2011], p. 1038), it is equally clear from the way in which *they* formalize their arguments that HRG, not HRS, is the version of Hamilton's rule *they* have in mind. Once we disambiguate the two superficially similar versions of Hamilton's rule, we are free to acknowledge both points: it is correct that HRS holds only in a very limited range of cases, but it is *also* correct that HRG holds extremely generally. This is why it is fitting to describe the two versions as, respectively, the 'special' and 'general' versions of Hamilton's rule.

3.1 Why HRS often fails

For a simple example of the limitations of HRS, we can return to the two-player, one-shot Prisoner's Dilemma and add a new payoff to the top-left box in the payoff matrix, denoted by

the letter *D* (see Queller [1984], [1985]; van Veelen [2009]; van Veelen et al. [2012]). *D* represents a *synergistic* payoff, a payoff that obtains only if both players cooperate:

	COOPERATE (Co_2)	**DEFECT** (De_2)
COOPERATE (Co_2)	$B - C + D$	$-C$
DEFECT (De_2)	B	0

D may be positive (perhaps representing a bonus payoff for cooperators that work together to achieve feats they could not achieve alone) or negative (perhaps representing diminishing returns caused by cooperators getting in each other's way). Either way, the size and magnitude of the *D*-payoff will plainly matter to the direction of evolution: for instance, if *D* is large and negative, cooperation could be much harder to evolve than it would be otherwise be; while, if *D* is large and positive, cooperation could evolve much more easily than it otherwise would. The upshot is that the condition for $W_{Co} > W_{De}$ is no longer given by $RB - C > 0$ (i.e., HRS), for this version of Hamilton's rule takes no account of the *D*-payoff. As Queller ([1984], [1985]) shows, the true condition for the evolution of cooperation in this game (henceforth: the 'synergy game') depends not merely on *R*, *B*, and *C*, but also on *D* and f_{Co}, the frequency of cooperators:[17]

[17] See (Queller [1984], [1985], [1992b], [2011]). The modified condition is sometimes known as 'Queller's rule' (see, e.g., Marshall [2011b]).

$$W_{\mathrm{Co}} > W_{\mathrm{De}} \text{ iff } RB - C + \left(\left(1 - R\right) f_{Co} + R\right) D > 0$$

Nowak, Tarnita and Wilson ([2010]) discuss a variety of more complicated cases in which HRS fails. It is hardly surprising, however, that HRS should fail in more complex cases, given that it fails even in very simple synergy games. In recent work, Matthijs van Veelen and colleagues (van Veelen [2009]; van Veelen et al. [2012]) show that the failure of HRS in the synergy game is merely one instance of a very general problem for HRS: HRS holds only if the difference between one's payoffs is independent of the strategy one's opponent plays, a condition sometimes called *Equal Gains from Switching (EGS)*:[18]

(EGS) (top left – bottom left) = (top right – bottom right)

When EGS holds, the f_{Co} and f_{De} terms cancel when one subtracts W_{De} from W_{Co}, leaving behind a simple inequality in terms of R, B and C. Crucially, however, EGS is likely to fail whenever there are non-additive fitness effects: that is, whenever an interaction between organisms produces a fitness effect that is more than a mere sum of the effects that each organism's individual behaviour, taken in isolation, would have had. There is no reason to suppose that non-additive effects of this nature are uncommon in nature. On the contrary, they are known to be widespread in insect societies (Anderson and McShea [2001]; Anderson

[18] For further discussion of Equal Gains from Switching, see (Nowak and Sigmund [1990]; Traulsen and Wild [2007]; van Veelen [2009]; van Veelen et al. [2012]).

and Franks [2001]; Anderson et al. [2001]) and in colonies of social microbes (Damore and Gore [2012]; Cornforth et al. [2012])). As a result, there is no reason to suppose that EGS applies generally or even particularly widely to real instances of social interaction.

3.2 Why HRG always holds

The same is *not* true of HRG. In recent work, Andy Gardner and colleagues (Gardner et al. [2007]; Gardner et al. [2011]) have shown that, when one construes Hamilton's rule as HRG, the rule still holds in games with fitness effects that depend non-additively on individual behaviours. In the case of the simple synergy game introduced above, all that happens is that the b and c coefficients in HRG come apart from the B and C payoffs in the payoff matrix, and now also depend on D, R, and f_{Co}:

$$b = B + \frac{1}{1+R}\left(R + \left(1-R\right) f_{\mathrm{Co}}\right) D$$

$$c = C - \frac{1}{1+R}\left(R + \left(1-R\right) f_{\mathrm{Co}}\right) D$$

HRG holds where HRS fails because, unlike HRS, it takes the D-payoff into consideration. It does so by means of a 'correction factor' in the cost and benefit coefficients: we account for synergy not as a third, separate predictor of the direction of evolution, distinct from c and b, but rather as a phenomenon that (if positive) lessens the average cost of cooperation and boosts its average benefit. The size of the correction factor depends on the

differential probability a cooperator has of receiving the synergistic payoff, and this probability is a function of *R*, the parameter that sets the differential probability of social partners playing identical strategies, and of f_{Co}, the overall frequency of cooperation in the population.

Although we should undoubtedly applaud Gardner and colleagues for showing exactly *how* HRG applies in the synergy game, it is important to see that HRG *cannot* fail to hold in any (real or modelled) system to which the Price equation applies, because it is simply an a priori implication of the Price equation and a linear regression model. Since the Price equation will still hold even when individuals interact in very complex ways, and since a linear regression model can be fitted to any set of population data[19], HRG will still hold in these cases. No matter how far we get from additive, pairwise interactions, we know that HRG cannot possibly fail unless the Price equation also fails.

4 A Dilemma for Hamilton's Defenders

It would be all too easy, at this point, to draw on the HRS/HRG distinction to accuse Nowak and colleagues of attacking a straw man. On the face of it, it seems that they are able to argue that Hamilton's rule 'almost never holds' only because they are construing Hamilton's rule in a particularly narrow sense (viz., as HRS). If they were to construe Hamilton's rule as many of their opponents do (viz., as HRG), then the basis for their criticisms would vanish.

[19] Subject to the proviso in footnote 13.

This, roughly speaking, is the response pressed by Gardner et al. ([2011]) (there is also a hint of this response in Abbot et al. [2011]). Nowak and his allies, however, maintain—not without some justification—that this response misunderstands the structure of their argument (see Nowak et al. [2011a]). For they reply that a retreat to HRG amounts to no more than a hollow victory for the kin selection theorist, on the basis that HRG is incapable of bearing the explanatory weight kin selection theorists expect it to carry:

> Hamilton's rule states that cooperation can evolve if relatedness exceeds the cost to benefit ratio. If cost and benefit are parameters of individual actions [HRS] then this rule almost never holds. There are attempts to make Hamilton's rule work by choosing generalized cost and benefit parameters [HRG], but these parameters are no longer properties of individual phenotypes. They depend on the entire system including population structure. These extended versions of Hamilton's rule have no explanatory power for theory or experiment. (Nowak et al. [2011a])

A related (though subtly different) complaint surfaces in Nowak, Tarnita and Wilson's online statement of June 2011:

> There exist generalized versions of Hamilton's rule that are designed to be 'always true' [HRG], but they are empty statements, which provide no insight for theory or experiment. (Nowak et al. [2011b])

Though one might be tempted to dismiss these replies as an *ad hoc* response to an unexpected barrage of criticism, this would not be fair: Nowak and his allies have been making arguments along these lines for some time. Indeed, Nowak, Tarnita and Wilson make a very similar argument against HRG in the supplementary material of their original article[20]:

> [W]hen realizing that the usual [$RB - C > 0$] rule [HRS] does not hold for a given model, Gardner et al (2007) propose that a modified rule [$rb - c > 0$] in fact holds, where [r] is [equal to] the usual relatedness but [b] and [c] are the 'effective' costs and benefits calculated using statistical methods [HRG] ... these effective costs and benefits unfortunately are very confusing and are typically functions of not only [B] and [C] but also of the relatedness R. Hence Hamilton's rule becomes [$R\mathrm{b}(R) - \mathrm{c}(R) > 0$], which makes it very complicated to separate any effects and it generally provides no intuition whatsoever. (Nowak et al. [2010], p. 18 (supplementary information))

It seems a little uncharitable, in light of these quotations, to accuse Nowak et al. of attacking a straw man. Rather, I suggest that Nowak, Tarnita and Wilson's case against Hamilton's rule is most charitably interpreted as presenting a dilemma for the kin selection theorist. The thought is that, however one prefers to interpret the 'cost', 'benefit' and 'relatedness' terms in

[20] I have made several alterations to the quotation to bring Nowak and colleagues' notation into line with my own. These do not affect the meaning of the quoted passage.

Hamilton's rule, the resultant principle is supposed to be both widely applicable and explanatorily powerful. Nowak and colleagues' argument is that:

(i) HRS is not widely applicable.

(ii) HRG has no explanatory power.

If they are right, then neither formulation satisfies both desiderata. Hamilton's rule—whichever formulation one prefers—is not the widely applicable explanatory principle that kin selection theorists take it to be.[21]

Do they have a case? For the reasons given in Section 3, the suggestion that HRS is not widely applicable seems entirely reasonable. But the claim that HRG buys its generality at the expense of explanatory power is more contentious—and Nowak and colleagues' argument for this claim is stated rather too briefly to be persuasive as it stands. They assert that HRG 'provides no intuition whatsoever', because the *b* and *c* coefficients 'are no longer properties of individual phenotypes' but instead 'depend on the entire system including population structure'; but the reasoning behind this assertion remains frustratingly opaque. In what sense are *b* and *c* 'no longer properties of individual phenotypes'? And in what sense does this result in them providing 'no intuition whatsoever'? In the remainder of the paper, I want to examine Nowak and colleagues' case against HRG in greater detail. To this end, it will be

[21] One possible response to Nowak and colleagues' 'dilemma' would be to argue that a third version of Hamilton's rule, distinct from both HRS and HRG, can satisfy both desiderata. Though I leave open the possibility of a 'third way' between HRS and HRG, I do not think any extant variants of Hamilton's rule can do this job. For instance, Hamilton's original ([1964]) route to Hamilton's rule—not discussed here—invokes similar assumptions to HRS, including the assumptions of weak selection and additive fitness effects. Many other derivations require similar assumptions.

helpful to disentangle two separate strands of criticism that have surfaced in the recent debate, and which Nowak et al. run together in the above quotations:

a) HRG has no explanatory power because it is 'always true'.

b) HRG has no explanatory power because the cost and benefit terms do not represent 'properties of individual phenotypes'.

These criticisms are quite distinct. Both, however, are somewhat obscure at first glance; and both must be explicated with care before we can see how the defender of HRG might respond. In the next two sections, I consider each in turn.

5 HRG and Explanatory Power I: The 'Tautology Problem' Redux

On several occasions, Nowak and his allies have expressed the concern that the derivation of HRG is a kind of black magic—that the rule appears from nowhere, pulled out of the Price equation like a rabbit from a hat.[22] This concern is not unreasonable. The Price equation is, after all, a highly abstract mathematical theorem which makes very few assumptions about the population it describes, and which tells us nothing at all about the conditions under which a social behaviour will be favoured by natural selection other than that it will be favoured

[22] See, e.g., (Nowak and Highfield [2011]): 'answers do indeed seem to pop out of the equation, like rabbits from a magician's hat' (p. 101); 'I found that the mathematical methods of kin selection were often murky. ... Equations seemed to arise out of nowhere in kin selection' (p. 104). As Nowak and Highfield note, the suggestion that results emerge from Price's formalism 'like rabbits from a hat' was first made by Hamilton himself, though not in any disparaging sense (see Hamilton [1996], p. 172).

when its breeding value co-varies positively with fitness.[23] For this reason, the equation is often described as a 'mathematical tautology', even by its foremost proponents (see, e.g., Frank [1995], [2012]; Okasha [2006]). To get from the Price equation to HRG, we need only substitute in a regression model which, owing to the residual term, is compatible with any possible set of population data. In effect, therefore, we start with *two* mathematical tautologies; and yet, by substituting one into the other, we arrive at a principle that is afforded huge explanatory significance by both theorists and experimentalists. Nowak and colleagues' concern can thus be expressed as follows: if the ingredients from which it is derived are merely mathematical tautologies, how can Hamilton's rule nevertheless carry the explanatory weight it is expected to carry? How can it tell us anything *empirically informative* about the ecological conditions for social evolution, if the equations from which it is derived do not?

This complaint against HRG has echoes of something much older. In the early days of the field, philosophers of biology were vexed by the so-called 'tautology problem': the charge that evolutionary theory is explanatorily empty because the phrase 'survival of the fittest' is a tautology. The phrase claims that the 'fittest' organisms survive, but the 'fittest' are (supposedly) defined as the organisms that survive; the phrase therefore tells us nothing about the evolutionary process that we could not have grasped simply by understanding the concepts it contains. While the alleged problem stimulated important work on the nature of fitness (e.g., Mills and Beatty [1979], Rosenberg [1983]; Sober [1984]), it is, in hindsight, rather hard to take seriously, firstly because fitness is not normally defined in terms of survival alone, and secondly because the phrase 'survival of the fittest' carries no serious

[23] van Veelen et al. ([2012]) compare this to Johan Cruyff's famous tactical advice: to win a game of football, you need to score more goals than your opponent.

explanatory weight in evolutionary theory (Dawkins [1982]). The nature of Nowak and colleagues' first complaint against HRG is, in some ways, strikingly similar. The difference, of course, is that the stakes are much higher. For Hamilton's rule, unlike 'survival of the fittest', is undoubtedly expected to carry serious explanatory weight.

How, then, can the defender of HRG respond to this new 'tautology' complaint? It will be instructive to consider one natural response which does not succeed. This is to argue that, although HRG is indeed an a priori implication of the Price equation and a linear regression model, all modelling results are in some sense a priori—so the explanatory power of the rule cannot be undermined by its a priority unless all modelling results are similarly undermined. This response, reasonable as it may sound on first hearing, fails to grasp an important difference between HRG and most other important modelling results in theoretical biology. For while it is true that modelling work often generates results that are arguably a priori, these are usually *conditional* results of the form: *if some substantive modelling assumptions obtain, then this outcome follows*. It is never a priori that a model will succeed in describing any *actual* evolutionary process, because it is never a priori that any actual evolutionary process will satisfy the antecedent of the conditional.[24] HRG is unusual not in that it is a priori, but in that it is a priori true of any possible evolutionary change, whether real or modelled, by virtue of the definitions of the terms involved. This is the sense in which it is 'tautologous', or 'always true'; and it is this combination of a priority and *unconditional* descriptive content that sets HRG apart from conditional results obtained within concrete models of particular scenarios. The question at issue, then, is this: how can a claim that is a

[24] van Veelen et al. ([2012]) make a similar point. See also (Sober [2011], Lange and Rosenberg [2011]) for discussion of this issue.

priori true of any possible evolutionary change—regardless of the nature of the entities in question, and regardless of how they interact with one another—tell us anything empirically informative about the evolution of social behaviour? The concern is that, because HRG fits everything, it explains nothing.

For a more promising line of response, we can begin by making explicit a point that is, I suspect, often simply taken for granted in discussions of HRG. This is that, in at least some circumstances, the partial regression of *Y* on *X* may be interpretable as a measure of the *causal effect* of *X* on *Y*.[25] Figure 1 shows how a causal interpretation of this sort might apply to the regression model we used to derive HRG.

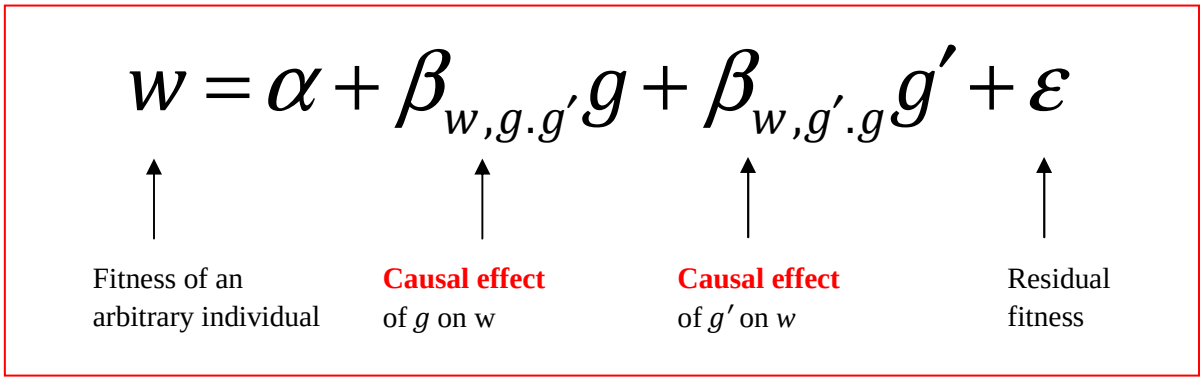

Figure 1. *The causal interpretation of a linear regression model.*

If a particular regression model admits of a causal interpretation, then substituting that equation into the Price equation allows us to decompose the overall *w*-*g* covariance into

[25] Why talk of 'causal effects' rather than simply 'effects'? The reason is that Fisher ([1930], [1941]) makes heavy use of the notion of the 'average effect' of a gene substitution on a phenotypic character, which he defines stipulatively as the partial regression of phenotypic value on allelic dosage, correcting for other alleles. Strictly speaking, this is a statistical notion rather than a causal one, so Fisher's choice of terminology is unfortunate. In talking of 'causal effects', I am referring specifically to partial regression coefficients *interpreted causally*.

separate causal components, each attributable to a distinct influence on fitness (see Frank [1998]; Okasha [2006]). In this way, partitioning the Price equation *adds causal content*: the partition entails claims about the causes of evolutionary change that the Price equation alone does not entail. Figure 2 shows the particular partition we used to derive HRG, together with its causal interpretation.

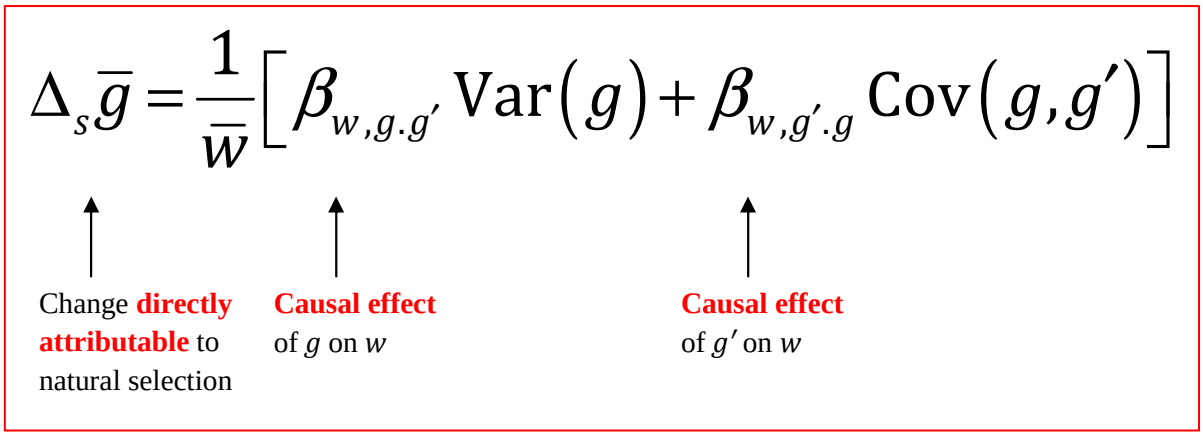

Figure 2. *The causal interpretation of the kin selection partition of the Price equation, derived in Section 2.2.*

By rearranging this expression following the procedure outlined in Section 2.2, while holding the causal interpretations of the relevant coefficients in place, we can derive a causal interpretation of HRG (Figure 3). Verbally, the causal interpretation of HRG states that selection will act to increase the population mean of g iff the causal effect of g on w, plus the relatedness-weighted causal effect of g' on w, exceeds zero. Thus interpreted, HRG embodies a substantive claim about the *causal* conditions for the evolution of social behaviour that the Price equation alone does not entail.

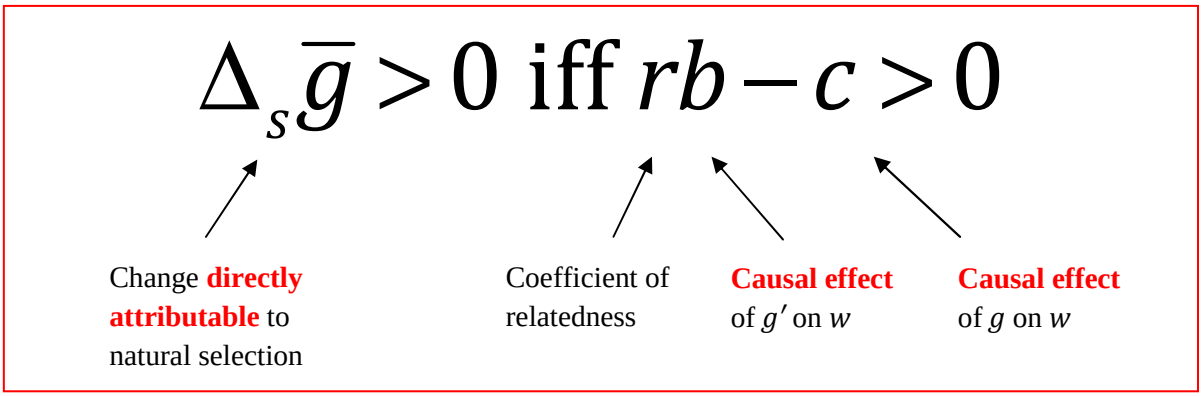

Figure 3. *The causal interpretation of HRG.*

To see how this bears on our 'tautology problem', it is crucial to note that regression coefficients are only *sometimes* causally interpretable. They are not *always* so interpretable: to think otherwise is to suppose that we can simply read off causation from measures of statistical association (Okasha [2006]).[26] Roughly speaking, the coefficients in a regression analysis of fitness will only admit of a causal interpretation if the predictors in the regression equation exhaustively account for the causal pathways through which fitness is influenced by genotype: if multiple pathways are conflated within the same term, or if some pathways are omitted altogether, regression coefficients will partly reflect spurious correlations rather than genuine causal influence (cf. Spirtes et al. [2000]). I revisit this point in Section 6.4, when I

[26] Note, for example, that coefficients of relatedness do not normally admit of a causal interpretation, because correlations between the genotypes of social partners are not normally due to the causal influence of one genotype over the other. Such correlation usually arises from a common cause that affects both genotypes (i.e., descent from a common ancestor).

consider in greater detail some of the conditions under which regression coefficients do and do not support a causal interpretation. For now, I merely want to draw attention to the fact that, since they are not always causally interpretable, a causal interpretation of a partition of the Price equation is not always true. The partition itself may hold for all possible populations, but the causal interpretation of that partition will not.

This allows us to see why the 'tautology' complaint misses its target. For I submit that, when theorists invoke HRG to do explanatory work, they are most charitably construed as invoking not merely HRG as a bare mathematical theorem, but rather the *interpretation* of that theorem in terms of causal effects. In Sections 2 and 3, we saw the importance of distinguishing between HRS and HRG in discussions of the explanatory scope of Hamilton's rule. It is no less important, I submit, to distinguish HRG-*qua-mathematical-theorem* from HRG-*qua-explanatory-principle*. When HRG is invoked qua explanatory principle, it is implicitly assumed that the *b* and *c* coefficients admit of a causal interpretation in terms of causal effects. The consequence is that HRG-qua-explanatory-principle is no tautology. For the *b* and *c* coefficients do not always admit of a causal interpretation, and we cannot know a priori whether, in any given context, the conditions for causal interpretability are satisfied (cf. Section 6.4).

6 HRG and Explanatory Power II: Prediction *versus* Unification

The tautology complaint misfires, then, because it does not distinguish HRG's mathematical representation from its causal interpretation. The causal interpretation carries the explanatory

weight, and it is not tautologous. As the quotations we considered in Section 3 make plain, however, this is not the only reservation Nowak and colleagues have about HRG. There is a further worry: the worry that HRG is explanatorily empty because the cost and benefit terms do not denote 'properties of individual phenotypes' (Nowak et al. [2011a]). In this section, I reconstruct and evaluate this second strand of criticism.

6.1 The predictive limitations of HRG

To understand Nowak and colleagues' second complaint, we can return to the synergy game introduced in Section 3. Recall that, when we add a *D*-payoff to the payoff matrix, HRG still holds, but the *b* and *c* regression coefficients are not simply functions of the *B* and *C payoffs* in the payoff matrix, but also functions of *D*, *R* and the frequency of cooperation, f_{Co}. This is a clear illustration of what Nowak et al. have in mind when they talk of the *b* and *c* coefficients turning out to depend, in many cases, on 'the entire system including population structure'. *R*, which sets the differential probability that social partners will play identical strategies, is an aspect of population structure; and f_{Co} is similarly a property of the population as a whole.

There can therefore be no doubt that, by defining the *b* and *c* terms in Hamilton's rule as partial regression coefficients—and thereby allowing them to float free of the *B* and *C* payoffs in the payoff matrix—we achieve a high level of generality in return for some potentially rather complicated and counterintuitive relationships between the coefficients in Hamilton's rule and the parameters of the model—or real population—to which we are

applying it. Gardner et al. ([2007]) openly concede that this is the price we have to pay for generality; but they evidently believe it to be a price worth paying. Here, they are plainly at odds with Nowak et al. ([2010]), for whom the counterintuitive relationships between the terms in HRG and the underlying parameters strip HRG of all explanatory value.

Who is correct here? One can certainly make a reasonable argument that, when *b* and *c* depend on population structure, HRG fails to license the sort of predictions one might intuitively expect Hamilton's rule to license. For example, one might expect Hamilton's rule to license the prediction that, in any given model, intervening on the parameters in such a way as to increase the relatedness, *r*, between actors and recipients would make the evolution of cooperation more likely.[27] In fact, HRG underwrites no such prediction. This is because, when the *r*, *b* and *c* coefficients in HRG are all functions of the same parameter[28], it is quite possible that intervening to increase *r* will make the evolution of cooperation *less* likely, , as our intervention may also have the effect of increasing *c* and decreasing *b*. It is even possible that $rb-c$ will be greater than zero prior to the intervention to increase *r,* yet less than zero afterwards, as a direct result of our intervention.[29]

---

[27] Note that, because *r* is a statistical property of the entire population, it is not a variable on which one can typically intervene *directly* in the context of formal modelling. Instead, one would intervene on the parameters of the model (such as the *R*-parameter in the synergy game) that determine the value of *r*. Nevertheless, one might intuitively expect it to be the case that intervening on the parameters that determine *r*, such that *r* increases, would promote the evolution of cooperation. It is important to see that this intuition is often incorrect.

[28] For example: in the synergy game, they are all functions of *R*, the parameter that sets the differential probability that social partners play identical strategies.

[29] This will occur in cases in which the *D*-payoff is negative and larger in magnitude than the *B*-payoff. For example, suppose $B=4$, $C=-2$, $D=-8$, and $f_{\text{Co}}=0.1$. At $R=0.25$, $rb-c=0.4$ and HRG is satisfied. But at $R=0.5$, $rb-c=-0.4$ and HRG is not satisfied. We can picture this as a case of *severely* diminishing returns: when cooperators interact with other cooperators, the diminishing returns *more than cancel out* the original

One might also expect it to be the case that, if HRG is satisfied, cooperation will tend to be selected in the long run. Again, however, HRG licences no such prediction. This is because, when the *b* and *c* coefficients are functions of f_{Co}, they are likely to change from one generation to the next as the frequency of cooperators changes. The implication is that the fact that HRG is satisfied in one generation cannot give us any confidence that it will still be satisfied in later generations, and therefore cannot give us any confidence that cooperation will be stable in the long run.[30]

It may be tempting, at this point, to reply that we *can* use HRG to predict the effects of interventions (and to predict long-run outcomes) as long as we understand the precise nature of the relationship between the coefficients in HRG and the parameters governing the evolutionary dynamics of the system under investigation. For instance, if we know the precise *way* in which intervening to increase *r* will impact on *b* and *c*, we will be able to predict whether this intervention will make the evolution of cooperation more or less likely. Yet this reply, it seems to me, does little to assuage our concerns about the predictive power of HRG. For the relationship between the terms in HRG and the dynamical parameters is highly *system-specific*: the results Gardner et al. ([2011]) derive in the context of two-player synergy games do not generalize to (say) three-player games, asymmetric games, *n*-player public goods games, and so on. If we accept that HRG is predictively inert in the absence of

---

benefit. Naturally, it is an empirical question how often scenarios of this sort arise in nature, but there is no reason to assume that they represent a merely theoretical possibility.

[30] The sign of $rb - c$ may be affected by changes in frequency whenever expected *D*-payoff makes a significant difference to the direction of evolution, since the expected *D*-payoff is sensitive to f_{Co}. For example, suppose again that $B = 4$, $C = -2$, $D = -8$, and $R = 0.25$. At $f_{\mathrm{Co}} = 0.1$, $rb - c = 0.4$ and HRG is satisfied. But at $f_{\mathrm{Co}} = 0.5$, $rb - c = -2$ and HRG is not satisfied.

system-specific functions relating its coefficients to the underlying dynamical parameters, we are, in effect, conceding that the predictive generality HRG appears to afford is illusory: to do serious predictive work with HRG, we must first find a way of expressing the relevant regression coefficients in terms of the dynamical parameters of the particular evolving system under study—and then the generality vanishes.

In summary, the problem for HRG can be put like this: while HRG holds for every evolving system to which the Price equation applies, it does not by itself entail any substantive predictions about the effects of interventions on these systems, or about how they will evolve in the long run. We can derive such predictions by *augmenting* HRG with expressions relating its terms to the dynamical parameters of the system under investigation, but these expressions are highly system-specific. We therefore face a trade-off. By construing Hamilton's rule as HRG rather than HRS, we buy generality at the expense of predictive power. We can buy back some of that predictive power by augmenting HRG with system-specific functions relating its coefficients to the underlying dynamics, but we do so at the expense of the predictive generality we originally hoped to gain.

6.2 The unification response

It therefore seems clear that HRG has serious predictive limitations. Yet whether that strips HRG of any explanatory power is another matter. Implicit in Nowak and colleagues' move from the predictive limitations of HRG to an attack on its 'explanatory power' is the assumption that, at least in the context of evolutionary theory, 'explanatory work' consists in

laying bare the dynamics of an evolutionary process in a way that enables long-run predictive success. The argument is that HRG, couched as it is in terms of overall statistical properties of the population, is no good for this kind of work—particularly in cases of non-additive interaction or frequency-dependent selection.

This is a conception of 'explanatory power' to which evolutionary game theory seems particularly well-suited. Yet it is not the *only* conception of explanatory power one might have. According to one long-running tradition in the philosophy of science, *unification* also counts for something: the thought, roughly speaking, is that, by bringing together disparate processes within a unifying framework, we increase our understanding of the causal structure of the world (see Kitcher [1989]; Strevens [2004], [2008]). I will not attempt to defend this conception of explanation here. I merely want to note that such a conception affords considerable value to unifying principles which abstract away from the details of particular models in order to capture, at a coarse-grained level, salient similarities between otherwise disparate processes. A unificationist conception of explanation allows that such generality has *intrinsic* explanatory value, whether or not it facilitates long-run predictive success or enables us to answer 'what-if-things-had-been-different' questions regarding the effects of interventions.

The defender of HRG can therefore allow that it has serious predictive limitations and yet maintain that it still has explanatory value. For HRG identifies a *common feature* that all processes of social evolution by natural selection must share: they are all processes that satisfy the condition $rb - c > 0$, where the coefficients *r*, *b* and *c* are understood in statistical

terms. In this way, HRG constitutes a unifying principle: a means of bringing together results from disparate models under a single conceptual framework.[31]

6.3 A worry about this response

HRG constitutes a very general condition that all processes of social evolution by natural selection must satisfy, regardless of their underlying causal differences; in this sense, it constitutes a unifying principle for social evolution theory. This seems like a promising response to Nowak and colleagues' criticisms of HRG, but it is not without its difficulties. One concern is that, if all we want is a condition that all processes of social evolution by natural selection must satisfy, we could achieve it rather more straightforwardly through 'Robertson's rule', which states that a social behaviour will be favoured by selection if and only if the simple regression of fitness on the genetic value for that behaviour is positive (see Robertson [1966]):

$$\textbf{(Robertson's rule)} \quad \Delta_s \bar{g} > 0 \text{ iff } \beta_{w,g} > 0$$

[31] Some remarks of Gardner et al. ([2007]) point towards a unificationist conception of the value of Hamilton's rule: 'The most powerful and simple approach to evolutionary problems is to start with a method such as population genetics, ... game theory or direct-fitness maximization techniques. The results of these analyses can then be interpreted within the frameworks that Price's theorem and Hamilton's rule provide. The correct use of these powerful theorems is to translate the results of such disparate analyses, conducted with a variety of methodologies and looking at very different problems, into the common language of social evolution theory' (Gardner et al. [2007], p. 224).

The derivation of Robertson's rule exactly parallels that of HRG (see Section 2.2). The only difference is that, instead of using two predictor variables in the linear regression model, we use a single predictor variable: the genetic value of the focal individual. This simple regression runs together the effects of the focal individual's genes and the effects of any correlated genes into a single measure of the overall association between genetic value and fitness. It is no less *accurate* for doing so: just like HRG, Robertson's rule is an a priori implication of the Price equation, and will hold in *any* population to which the Price equation applies. Consequently, it too identifies a common feature that all processes of social evolution by natural selection must share: they are all processes for which $\beta_{w,g} > 0$.

It is tempting to object that, while Robertson's rule may be fully general, it is just too simple to be predictively useful when organisms interact socially, for it compresses all the causal influences on the direction of evolution into a single regression coefficient. Yet, while this is true enough, we have already seen that HRG *also* compresses the causal influences on the direction of evolution into a small number of coefficients in a way that impairs *its* predictive utility whenever social interactions are non-additive. The only difference is that HRG uses *two* coefficients rather than one: it partitions the simple regression of fitness on genetic value into a '$-c$' component and an 'rb' component.

The worry is that, if *all* we care about is generality, then nothing is gained by partitioning the simple regression of fitness on genotype into two components: we may as well use Robertson's rule as our unifying principle. But if we want a rule that represents each of the distinct causal influences on fitness in a separate term, and that is therefore useful for prediction, intervention and causal explanation, we will often need to split the regression of

fitness on genotype into *more* than two components (cf. Queller [1985], [1992b], [2011]; Frank [1998]). Either way, the theoretical role for HRG appears to be somewhat limited.

6.4 Causal interpretation revisited

In Section 5, we encountered the suggestion that, at least under some circumstances, regression coefficients may be interpreted as measures of causal effects, and we saw how distinguishing the mathematical representation of Hamilton's rule from its causal interpretation exposes a flaw in the 'tautology' complaint against HRG. Now, I want to suggest that the causal interpretability of regression coefficients also shows how HRG can perform a theoretical function that Robertson's rule cannot. For, crucially, *HRG admits of a causal interpretation under a broader range of conditions than Robertson's rule.* As a result, HRG, in contrast to Robertson's rule, identifies an important *causal* feature that unites the processes by which social behaviour evolves.

To see why HRG admits of a causal interpretation under a broader range of conditions than Robertson's rule, we can return to the synergy game. Since Robertson's rule, like HRG, holds for *any* evolutionary process (by virtue of being an a priori implication of the Price equation), it holds in the synergy game. Its single coefficient, $\beta_{w,g}$, compresses all the effects relevant to the direction of evolution—namely, the B, C, and D payoffs—into a single, overall measure of the statistical association between fitness and genetic value:

$$\beta_{w,g} = -C + RB + \left(R + \left(1-R\right) f_{\mathrm{Co}}\right) D$$

In compressing all the influences on the direction of evolution into a single coefficient, $\beta_{w,g}$ conflates distinct causal pathways; the result is that it cannot plausibly be regarded as measuring the causal effect of one's own genotype on one's fitness. One way to see this is to imagine what we would have to say, if we insisted on interpreting it in this way: we would have to say that the focal individual is causally responsible not merely for C, a payoff that results directly from its own behaviour, but also for B, a payoff that results directly from the behaviour of its social partners. Plainly, this interpretation could only be correct if the focal individual were causally responsible for the behaviour of its social partners; and, while this might be true in some cases, it is not true in the synergy game.[32] This is just one instance of a quite general problem for the causal interpretability of Robertson's rule: if $\beta_{w,g}$ is to sustain a causal interpretation, one's own genotype must be causally responsible for all the fitness effects with which it correlates. Provided each social partner retains control of its own behaviour, this assumption will fail in any case in which genetic relatives interact socially.

The situation for the regression coefficients in HRG is not so bleak. Recall that, in a synergy game, the b and c regression coefficients in HRG take account not only of the B and C payoffs in the payoff matrix, but also of the D payoff. They do so in a way that splits the expected synergistic effect evenly between the b and c terms:

[32] If R is positive, the behaviour of the focal individual will correlate with the behaviour of its social partners, and this will lead to a *non-causal* correlation between its own behaviour and the probability that it receives the B payoff. There is no suggestion, however, that the behaviour of its social partners causally depends on its own behaviour, and this is what would have to be the case for $\beta_{w,g}$ to sustain a causal interpretation.

$$b = B + \frac{1}{1+R}\left(R + \left(1-R\right) f_{\mathrm{Co}}\right) D$$

$$c = C - \frac{1}{1+R}\left(R + \left(1-R\right) f_{\mathrm{Co}}\right) D$$

Can *b* and *c* be interpreted as measures of causal effects in the synergy game? They can, conditional on the assumption that, when a fitness effect depends symmetrically on the behaviour of two agents, it is reasonable to attribute an equal portion of the resultant effect to the behaviour of each agent. If this assumption is correct, then *b* and *c* do indeed measure, respectively, the causal effect of one's own genotype on one's fitness and of one's social partner's genotype on one's fitness, since they correctly apportion causal responsibility for both the linear effects *B* and *C* and the synergistic effect *D*. Moreover, I submit that the assumption is at least *prima facie* plausible. Critics of HRG may wish to dispute this assumption, and this may present a productive avenue for further discussion. For now, I merely want to note that, conditional on a plausible assumption about how causal responsibility ought to be apportioned in cases of synergistic interaction, the coefficients in HRG admit of a causal interpretation in such cases, even though they cannot be equated with any of the parameters in the payoff matrix.

This suggests that our concern about the unification response was misplaced. It is true enough that HRG is not the simplest rule one can formulate regarding the conditions under which a social behaviour is favoured by natural selection. But it is the simplest such rule that also plausibly admits of a *causal interpretation* across a wide range of cases, including cases

of synergistic interaction. The upshot is that HRG-qua-explanatory-principle captures a substantial causal insight about the evolution of social behaviour that Robertson's rule does not capture. This is the insight that many, perhaps all, of the processes through which social behaviour evolves are united by the following causal feature: they are processes in which the causal effect of an individual's genotype on its own fitness, plus the relatedness-scaled causal effect of its social partner's genotype on its fitness, is greater than zero.

This point brings together the discussions in Sections 5 and 6. The causal interpretability of regression coefficients under appropriate conditions shows how HRG can be more than a mere tautology. But it does something else too: it also shows how HRG can serve an important explanatory function irrespective of whether it enables long-run predictive success. Because its coefficients are causally interpretable in a wide range of cases, HRG-qua-explanatory-principle identifies a substantial *causal* unity to the processes by which social behaviour evolves. Because its coefficients are not causally interpretable in a wide range of cases, Robertson's rule does not.

7 The Heart of the Matter

The current controversy regarding Hamilton's rule has brought to the fore subtle but divisive issues in the foundations of social evolution theory. While the bones of contention are also partly empirical—and while Nowak et al. are surely guilty of underplaying kin selection's empirical track record—there are also significant conceptual issues at stake. I have argued that, to understand the nature of the debate, we need to distinguish two versions of

Hamilton's rule: a special version (HRS) in which the '*b*' and '*c*' terms represent fecundity payoffs; and a general version (HRG), derived from the Price equation, in which the '*b*' and '*c*' terms represent partial regression coefficients. And I have argued that, on a charitable reconstruction, Nowak and colleagues' argument is that HRS almost never holds, while HRG buys its generality at the expense of explanatory power. While their criticisms of HRS are difficult to argue with, their criticisms of HRG are more contentious. Yet they have gone largely unanswered in the subsequent debate.

Close examination of these criticisms reveals the importance of a further distinction: that between HRG-qua-mathematical-theorem and HRG-qua-explanatory-principle, where the latter takes for granted the causal interpretability of the '*b*' and '*c*' coefficients. This distinction, and the attention it draws to questions of causal interpretability, is valuable for two reasons. First, it shows how HRG can come to embody substantial causal content about social-evolutionary processes—content that is empirically grounded in a wide range of cases, but that is by no means 'always true'—even though its formal derivation makes no substantive assumptions about the population it describes. This helps assuage Nowak and colleagues' concern that HRG is no more than a mathematical tautology. Second, the distinction shows how HRG can serve a valuable explanatory function in spite of its serious predictive limitations. For the primary theoretical value of HRG lies not in its ability to underwrite predictions about long-run evolutionary outcomes, but rather in its identification of a common causal feature that unites the processes by which social behaviour evolves.

The considerations I have brought to bear are not intended to settle the debate once and for all. Live issues remain—in particular, issues concerning the relative importance of

unification and prediction in evolutionary explanations, and concerning the correct procedure for apportioning causal responsibility in cases of synergistic interaction—that may yet divide Hamilton's defenders from their opponents, and that are unlikely to be settled definitively by empirical or theoretical considerations alone. The debate is therefore unlikely to go away. My primary aim has been to clarify precisely what is at stake, and to give Hamilton's defenders and opponents a common vocabulary in which to communicate with one another. Throughout much of the debate following Nowak and colleagues' ([2010]) article, theorists have been talking at cross-purposes. Even the notion at the very heart of the debate—'Hamilton's rule'—is ambiguous. Only by distinguishing special and general versions of the rule, and by distinguishing the rule's mathematical representation from its causal interpretation, can we hope to move towards a more productive discussion of its uses and limits.

Funding

Christ's College; Arts and Humanities Research Council.

Acknowledgements

Versions of this paper were presented at the Serious Metaphysics Group, University of Cambridge, in November 2011; and at Philosophy of Biology in the UK, All Souls College, Oxford, in April 2012. I am grateful to the audiences at these events for extremely helpful questions and comments. For detailed comments on the manuscript, I thank Tim Button, Ellen Clarke, Sarah Coakley, Herbert Gintis, Bram Kuijper, Tim Lewens, Johannes Martens,

Bence Nanay, Robert Northcott, Martin Nowak, Samir Okasha, Cedric Paternotte and four anonymous referees.

Christ's College

St Andrew's Street

Cambridge, CB2 3RU, UK

jgb37@cam.ac.uk

References

Abbot, P., et al. [2011]: 'Inclusive Fitness Theory and Eusociality', *Nature*, **471**, pp. E1-E2.

Anderson C., Franks, N. R. and McShea, D. W. [2001]: 'The Complexity and Hierarchical Structure of Tasks in Insect Societies', *Animal Behaviour*, **62**, pp. 643-51.

Anderson, C. and Franks, N. R. [2001]: 'Teams in Animal Societies', *Behavioural Ecology*, **12**, pp. 534-40.

Anderson, C. and McShea, D. W. [2001]: 'Individual *versus* Social Complexity, with Particular Reference to Ant Colonies', *Biological Reviews*, **76**, pp. 211-37.

Boomsma, J. J., Beekman, M., Cornwallis, C. K., Griffin, A. S., Holman L., Hughes, W. O. H., Keller, L., Oldroyd, B. P. and Ratnieks, F. L. W. [2011]: 'Only Full-sibling Families Evolved Eusociality', *Nature*, **471**, pp. E4–E5.

Bourke, A. F. G. [2011a]: 'The Validity and Value of Inclusive Fitness Theory', *Proceedings of the Royal Society B*, **278**, pp. 3313-20.

Bourke, A. F. G. [2011b]: *Principles of Social Evolution*, Oxford: Oxford University Press.

Bourke, A. F. G. and Franks, N. R. [1995]: *Social Evolution in Ants*, Princeton, NJ: Princeton University Press.

Charlesworth, B. [1980]: 'Models of Kin Selection', in H. Markl (ed.), 1980, *Evolution of Social Behaviour: Hypotheses and Empirical Tests*, Weinheim: Verlag, pp. 11-26.

Charnov, E. L. [1977]: 'An Elementary Treatment of the Genetical Theory of Kin Selection', *Journal of Theoretical Biology*, **66**, pp. 541–50.

Cornforth, D. M., Sumpter, D. J., Brown, S. P. and Brännström, A. [2012]: 'Synergy and Group Size in Microbial Cooperation', *American Naturalist*, **180**, pp. 296-305.

Damore, J. A., and Gore, J. [2012]: 'Understanding Microbial Cooperation', *Journal of Theoretical Biology*, **299**, pp. 31-41.

Damuth, J. and Heisler, I. L. [1988]: 'Alternative Formulations of Multilevel Selection', *Biology and Philosophy*, **3**, pp. 407–30.

Dawkins, R. [1982]: *The Extended Phenotype: The Gene as the Unit of Selection*, New York: W. H. Freeman.

Doebeli, M. [2010]: 'Inclusive Fitness is Just Bookkeeping', *Nature*, **467**, p. 7316.

Falconer, D. S. and Mackay, T. F. C. [1996]: *Introduction to Quantitative Genetics* (4th edition), London: Longman.

Ferriere, R. and Michod, R. E. [2011]: 'Inclusive Fitness in Evolution', *Nature*, **471**, pp. E6-E7.

Fisher, R. A. [1930]: *The Genetical Theory of Natural Selection*, Oxford: Clarendon Press.

Fisher, R. A. [1941]: 'Average Excess and Average Effect of a Gene Substitution', *Annals of Human Genetics*, **11**, pp. 53-63.

Frank, S. A. [1995]: 'George Price's Contributions to Evolutionary Genetics', *Journal of Theoretical Biology*, 175, pp. 373-88.

Frank, S. A. [1998]: *Foundations of Social Evolution*, Princeton, NJ: Princeton University Press.

Frank, S. A. [2012]: 'Natural Selection. IV. The Price Equation', *Journal of Evolutionary Biology*, **25**, pp. 1002-19.

Gardner, A. [2008]: 'The Price Equation', *Current Biology*, **18**, pp. R198-R202.

Gardner, A. and Foster, K. R. [2008]: 'The Evolution and Ecology of Cooperation—History and Concepts', in J. Korb and J. Heinze (eds), *Ecology of Social Evolution*, 2008, Heidelberg: Springer-Verlag, pp. 1-36.

Gardner, A., West, S. A. and Barton, N. H. [2007]: 'The Relation between Multilocus Population Genetics and Social Evolution Theory', *American Naturalist*, **169**, pp. 207-26.

Gardner, A., West, S. A. and Wild, G. [2011]: 'The Genetical Theory of Kin Selection', *Journal of Evolutionary Biology*, **24**, pp. 1020-43.

Grafen, A. [1985]: 'A Geometrical View of Relatedness', *Oxford Surveys in Evolutionary Biology*, **2**, pp. 28-89.

Grafen, A. [2006]: 'Optimization of Inclusive Fitness', *Journal of Theoretical Biology*, **238**, pp. 541-63.

Hamilton, W. D. [1964]: 'The Genetical Evolution of Social Behaviour', *Journal of Theoretical Biology*, **7**, pp. 1-52.

Hamilton, W. D. [1970]: 'Selfish and Spiteful Behaviour in an Evolutionary Model', *Nature*, **228**, pp. 1218-20.

Hamilton, W. D. [1972]: 'Altruism and Related Phenomena, Mainly in Social Insects', *Annual Review of Ecology and Systematics*, **3**, pp. 193-232.

Hamilton, W. D. [1975]: 'Innate Social Aptitudes of Man: An Approach from Evolutionary Genetics', in R. Fox (ed.), *Biosocial Anthropology*, New York: Wiley, pp. 133-55.

Hamilton, W. D. [1996]: *Narrow Roads of Gene Land: The Collected Papers of W. D. Hamilton Volume 1: Evolution of Social Behaviour*, New York: W. H. Freeman.

Heisler, I. L. and Damuth, J. [1987]: 'A Method for Analyzing Selection in Hierarchically Structured Populations', *American Naturalist*, **130**, pp. 582–602.

Herre, E. A. and Wcislo, W. T. [2011]: 'In Defence of Inclusive Fitness Theory', *Nature*, **471**, p. E8.

Kitcher, P. [1989]: 'Explanatory Unification and the Causal Structure of the World', in P. Kitcher and W. C. Salmon (eds), 1989, *Scientific Explanation*, Minneapolis: University of Minnesota Press, pp. 410-505.

Lande, R., and Arnold, S. J. [1983]: 'The Measurement of Selection on Correlated Characters', *Evolution*, **37**, pp. 1210-26.

Lange, M., and Rosenberg, A. [2011]: 'Can There Be *A Priori* Causal Models of Natural Selection?' *Australasian Journal of Philosophy*, **89**, pp. 591-9.

Marshall, J. A. R. [2011a]: 'Group Selection and Kin Selection: Formally Equivalent Approaches', *Trends in Ecology and Evolution*, **26**, pp. 325-32.

Marshall, J. A. R. [2011b]: 'Queller's Rule OK: Comment on van Veelen, "When Inclusive Fitness Is Right and When It Can Be Wrong"', *Journal of Theoretical Biology*, **270**, 185-8.

Michod, R. E. [1982]: 'The Theory of Kin Selection', *Annual Review of Ecology and Systematics*, **13**, pp. 23-55.

Mills, S. K., and Beatty, J. H. [1979]: 'The Propensity Interpretation of Fitness', *Philosophy of Science*, **46**, pp. 263-86.

Mueller, L. D. and Feldman, M. W. [1985]: 'Population Genetic Theory of Kin Selection: A Two-Locus Model', *American Naturalist*, **125**, pp. 535-49.

Nowak, M. A. [2006]: 'Five Rules for the Evolution of Cooperation', *Science*, **314**, pp. 1560-3.

Nowak, M. A. and Highfield, R. [2011]: *SuperCooperators: Evolution, Altruism, and Why We Need Each Other to Succeed*, New York: Free Press.

Nowak, M. A. and Sigmund, K. [1990]: 'The Evolution of Stochastic Strategies in the Prisoner's Dilemma", *Acta Applicandae Mathematicae*, **20**, 247-65.

Nowak, M. A., Tarnita, C. E. and Wilson, E. O. [2010]: 'The Evolution of Eusociality', *Nature*, **466**, pp. 1057-62.

Nowak, M. A., Tarnita, C. E. and Wilson, E. O. [2011a]: 'Nowak et al. reply', *Nature*, **471**, pp. E9-E10.

Nowak, M. A., Tarnita, C. E. and Wilson, E. O. [2011b]: 'A Brief Statement about Inclusive Fitness and Eusociality', <www.ped.fas.harvard.edu/IF_Statement.pdf>

Okasha, S. [2006]: *Evolution and the Levels of Selection*, Oxford: Oxford University Press.

Okasha, S. [2008]: 'Biological Altruism', in E. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2008 edition), <plato.stanford.edu>

Orlove, M. J. [1975]: 'A Model of Kin Selection Not Invoking Coefficients of Relationship,' *Journal of Theoretical Biology*, **49**, pp. 289–310.

Price, G. R. [1970]: 'Selection and Covariance', *Nature*, **227**, pp. 520-1.

Price, G. R. [1972]: 'Extension of Covariance Selection Mathematics', *Annals of Human Genetics*, **35**, pp. 485-90.

Queller, D. C. [1984]: 'Kin Selection and Frequency Dependence: A Game-Theoretic Approach', *Biological Journal of the Linnaean Society*, **23**, pp. 133-43.

Queller, D. C. [1985]: 'Kinship, Reciprocity, and Synergism in the Evolution of Social Behaviour', *Nature*, **318**, pp. 366-7.

Queller, D. C. [1992a]: 'A General Model for Kin Selection', *Evolution*, **46**, pp. 376-80.

Queller, D. C. [1992b]: 'Quantitative Genetics, Inclusive Fitness and Group Selection', *American Naturalist*, **139**, pp. 540-58.

Queller, D. C. [2011]: 'Expanded Social Fitness and Hamilton's Rule for Kin, Kith and Kind', *Proceedings of the National Academy of Sciences USA*, **108**, pp. 10792-9.

Robertson, A. [1966]: 'A Mathematical Model of the Culling Process in Dairy Cattle', *Animal Production*, **8**, pp. 95-108.

Rosenberg, A. [1983]: 'Fitness', *Journal of Philosophy*, **80**, pp. 457-73.

Rice, S. [2004]: *Evolutionary Theory: Mathematical and Conceptual Foundations*, Sunderland, MA: Sinauer.

Rousset, F., and Lion, S. [2011]: 'Much Ado about Nothing: Nowak et al.'s Charge against Inclusive Fitness Theory', *Journal of Evolutionary Biology*, **24**, pp. 1386-92.

Skyrms, B. [1996]: *The Evolution of the Social Contract*, Cambridge: Cambridge University Press.

Sober, E. [1984]: *The Nature of Selection: Evolutionary Theory in Philosophical Focus*, Chicago, IL: University of Chicago Press.

Sober, E. [2011]: '*A Priori* Causal Models of Natural Selection', *Australasian Journal of Philosophy*, **89**, pp. 571-89.

Spirtes, P., Glymour, C. and Scheines, R. [2000]: *Causation, Prediction and Search* (2nd edition), Cambridge, MA: MIT Press.

Strassmann, J. E., Page, R. E. Jr, Robinson, G. E. and Seeley, T. D. [2011]: 'Kin Selection and Eusociality', *Nature*, **471**, pp. E5-E6.

Strevens, M. [2004]: 'The Causal and Unification Approaches to Explanation Unified—Causally', *Noûs*, **38**, pp. 154-76.

Strevens, M. [2008]: *Depth: An Account of Scientific Explanation*, Cambridge, MA: Harvard University Press.

Taylor, C. and Nowak, M. A. [2007]: 'Transforming the Dilemma', *Evolution*, **61**, pp. 2281-92.

Toro, M., Abugov, R., Charlesworth, B. and Michod, R. E. [1982]: 'Exact *versus* Heuristic Models of Kin Selection', *Journal of Theoretical Biology*, **97**, pp. 699-713.

Trivers, R. L. [1985]: *Social Evolution*, Menlo Park, CA: Benjamin/Cummings.

Uyenoyama, M. K. and Feldman, M. W. [1980]: 'Theories of Kin and Group Selection: A Population Genetics Perspective', *Theoretical Population Biology*, **17**, pp. 380-414.

Uyenoyama, M. K. and Feldman, M. W. [1981]: 'On Relatedness and Adaptive Topography in Kin Selection', *Theoretical Population Biology*, **19**, pp. 87-123.

Uyenoyama, M. K. and Feldman, M. W. [1982]: 'Population Genetic Theory of Kin Selection II: The Multiplicative Model', *American Naturalist*, **120**, pp. 614-27.

Uyenoyama, M. K., Feldman, M. W. and Mueller, L. D. [1981]: 'Population Genetic Theory of Kin Selection: Multiple Alleles at One Locus', *Proceedings of the National Academy of Sciences USA*, **78**, pp. 5036-40.

van Veelen, M. [2005]: 'On the Use of the Price Equation', *Journal of Theoretical Biology*, **237**, pp. 412-26.

van Veelen, M. [2009]: 'Group Selection, Kin Selection, Altruism, and Cooperation: When Inclusive Fitness is Right and When It Can Be Wrong', *Journal of Theoretical Biology*, **259**, pp. 589-600.

van Veelen, M., Garcia, J., Sabelis, M. W. and Egas, M. [2010]: 'Call for a Return to Rigour in Models', *Nature*, **467**, p. 661.

van Veelen, M., Garcia, J., Sabelis, M. W., and Egas, M. [2012]: 'Group Selection and Inclusive Fitness are Not Equivalent; the Price Equation vs Models and Statistics', *Journal of Theoretical Biology*, **299**, pp. 64-80.

Wenseleers, T., Gardner, A. and Foster, K. R. [2010]: 'Social Evolution Theory: A Review of Methods and Approaches', in T. Székely, A. J. Moore and J. Komdeur (eds), 2010, *Social Behaviour: Genes, Ecology and Evolution*, Cambridge: Cambridge University Press, pp. 132-58.

West, S. A., Griffin, A. S. and Gardner, A. [2007]: 'Social Semantics: Altruism, Cooperation, Mutualism, Strong Reciprocity and Group Selection', *Journal of Evolutionary Biology*, **20**, pp. 415-32.

Wild, G., and Traulsen, A. [2007]: 'The Different Limits of Weak Selection and the Evolutionary Dynamics of Finite Populations', *Journal of Theoretical Biology*, **247**, pp. 382-90.