TI 2024-033/I
Tinbergen Institute Discussion Paper

The generalized version of Hamilton's rule

Matthijs van Veelen[1]

[1] University of Amsterdam and Tinbergen Institute

The generalized version of Hamilton's rule

Abstract. The main ingredient of this paper is the derivation of the generalized version of Hamilton's rule. This version is derived with the Generalized Price equation. The generalized version of Hamilton's rule generalizes the original rule, in the sense that it produces a set of rules; one rule for every different model of how social interactions affect fitnesses. Every such Hamilton-like rule is generally valid; they all correctly determine when altruism, or costly cooperation, will be selected for, whatever model they are combined with. Every such rule, however, only has a meaningful interpretation in combination with the model it belongs to. The classic Hamilton's rule is the generalized Hamilton's rule that goes with the linear model. The insight that there are many Hamilton-like rules, all of which are generally valid, but none of which is generally meaningful, helps understand the controversy surrounding Hamilton's rule, and provides a constructive way to always find a rule that both gets the direction of selection right, and has a meaningful interpretation.

1. Introduction

Hamilton's rule is probably the most famous rule in evolutionary biology. The rule states that altruism will evolve if $rb > c$, where b is the fitness benefit to the recipient, c is the fitness cost to the donor, and r is the genetic relatedness between them. The paper in which it was presented has about half as many citations as Darwin's "*On the Origin of Species*" [27] and the rule is one of the core ingredients of "*The Selfish Gene*" by Richard Dawkins [28], which is the most popular of all popular science books. The generality of Hamilton's rule however has always been a topic of contention, and positions range all the way from "*Hamilton's rule almost never holds*" [29] to "*Inclusive fitness is as general as the genetical theory of natural selection itself*"[30]. This raises a range of questions. One of them, obviously, is how generally Hamilton's rule applies. Other questions concern the debate itself and include what explains the lack of convergence.

The crucial ingredient for answering all of these questions has to do with the Price equation. This equation was not used in the original derivation of Hamilton's rule [31]. The 1964 paper contains a model, and the results in it follow from its assumptions (see also [32], for a missing step in the derivation, and [24], for the relation with other well-known results at the time). Later versions do use the Price equation. Hamilton himself started using it in [33] and [34], and also later in the literature, the Price equation approach (a.k.a. the regression method) has been used for deriving Hamilton's rule. Examples include [7], [10-12], [15], [19], [20], [36]. Along with the Price equation came the claim of generality.

In the twin TI discussion paper on the Generalized Price equation, I show that what is missing in [1] is a few lines of essential algebra. Without these, the Price equation loses the link to normal statistics and modeling. With them, the link is restored, and the Generalized Price Equation is derived. This produces a Price-like equation for every (statistical) model, as long as it includes a constant and a linear term. All of these equations are general, in the sense that they are identities, also for all models other than the ones that they belong to, or for datasets that give us no reason to think they are generated by the statistical model they are formulated for. The terms in these Price equations however are only meaningful within the confines of their own model.

Both the shortcomings of the original Price equation, and the ways to mend them by using the generalized version, are mirrored in the shortcomings of Hamilton's rule as we know it, and ways to address those. Just like there is not just one Price equation, but a multitude of Price-like equations, there is not just one Hamilton's rule, but a multitude of Hamilton-like rules. All of them are correct, and all of them are general, but none of them is generally meaningful; all of them only have meaning under the model they are associated with.

In Section 2 we look at three models and three Hamilton-like rules. If the genetic setup is binary, in the sense that it features the presence or absence of a (possibly social) gene, then these three models exhaust all possibilities. The Hamilton-like rules for all three are derived using the Generalized Price equation. This follows Section 5 of the twin TI discussion paper, but with a social interpretation of the *p*-scores and the *q*-scores featuring there. (The notation here will also deviate a bit for clarity of exposition; instead of using β's for all coefficient in all models, we will use one Greek letter, with subscripts, per model). This illustrates the possibility that there can be multiple rules, all of which are general, in the sense that they all get the direction of selection right for all models, while the terms in them only have a meaningful interpretation in a restricted set of models. In other words, this illustrates that rules, like models, can be over- or underspecified.

In Section 3 I derive the Generalized Hamilton's rule, allowing for a richer genetic setup, where *p*-scores and *q*-scores are not restricted to be binary. This comes with a richer set of Hamilton-like rules implied by the Generalized Price equation.

2. Three models, three Price-like equations, three Hamilton-like rules

There are a few things that we keep simple. Reproduction is asexual. That means that we cannot think of pairs of interacting individuals as siblings or cousins, because that would require sexual reproduction. Hamilton's rule however is not confined by what causes relatedness between interacting individuals. The literature abounds with models with asexual selection and, for instance, local interaction structures, which induce relatedness (see for instance the overview by [36]). Asexual reproduction is easier to model, and what we find will carry over to models with sexual reproduction and kin recognition.

The classical setup with the Price equation assumes a parent population consisting of n individuals. Individual i is characterized by a *p*-score p_i and a *q*-score q_i. Something else that we keep simple, is that in this section, we assume that p_i is binary; it is 1 if a certain gene is present, and 0 otherwise. In Section 4 of the twin TI discussion paper on the Generalized Price equation, we have assumed that *q*-scores represent the presence or absence of a gene or a set of genes other than those represented by the *p*-score in person i herself. Here we switch to *q*-scores representing the presence or absence of the same gene that is represented by the *p*-score, but in the individual that individual i is interacting with. With sexual reproduction, we could think of this as a sibling or a cousin. With asexual reproduction this could for instance be an individual linked to individual i in homogeneous graph [37], or an individual within the same group [38]. The derivation of the Generalized

Price equation is not affected by the interpretation or the mode of reproduction (see the twin TI discussion paper on the Generalized Price equation).

The different models below represent different ways in which an individual's fitness can depend on its *p*-score and on its *q*-score (which is the *p*-score of its relative). In Model 1, the *q*-score does not affect individual i's fitness. In Models 2 and 3 it does, and there, the *p*-score of individual i will represent i's level of cooperation. In evolutionary game theory terms, $p_i = 0$ implies that individual i plays D, and $p_i = 1$ implies that individual i plays C.

The number of offspring that any individual can have must be an integer. A model would specify or imply a random variable over these integers, and the fitness of an individual is the expected value of this random variable. The models below will therefore include an error term to reflect the randomness. We will however assume that the population is sufficiently large for us to approximate it with an infinitely large population. In this infinite population, the number of offspring of any individual with a certain *p*-score and a certain *q*-score is still a random variable, but the average of all individuals with the same combination of *p*-score and *q*-score coincides with the expected value that the model specifies. Assuming an infinite population therefore allows us to apply the Generalized Price equation to the model, which, at the population level, has become deterministic. Not all papers in the literature are explicit or precise about this, but here we would like to not skip over this; see also Section 5 in the twin TI discussion paper on the Generalized Price equation.

The simple setup, with asexual reproduction, and binary *p*-scores and *q*-scores, limits the scope of possibilities to three models. This is however enough to illustrate that there can be different rules, all of which are equally valid, and equally general, while none of these rules, at the same time, is generally meaningful. They are meaningful for their own model but are over- or underspecified when applied to other models.

Model 1, Price equation 1, rule 1. The first is the linear model, in which the *p*-score of the individual only affects her own fitness, and nobody else's. Consequently, the *q*-score, or the *p*-score of the individual it interacts with, has no effect on her fitness:

$$w_i = \beta_0 + \beta_1 p_i + \varepsilon_i$$

In an infinite population, the Generalized Price equation for the first model is

$$\bar{w}\Delta\bar{p} = \hat{\beta}_1 \mathrm{Var}(p)$$

where $\hat{\beta}_1$ is independent of the composition of the parent population, and equal to β_1. This is straightforward, but just to be sure, we do this step-by-step in Appendix A.1, also to stress the role of the assumption of large populations.

From this, we can see that $\Delta\bar{p} > 0$ if and only if $\hat{\beta}_1 > 0$. This is clearly a rule, and it is so straightforward that everyone in evolutionary biology uses it for non-social traits with linear fitness effects, often without even referring to it as a rule. The coefficient $\hat{\beta}_1$ has a meaningful interpretation; it is the effect of the gene on its carrier, and absent any effect on or from interaction partners, this is all that matters for whether this gene will be selected.

Later, when we compare the different models and the different rules, it may help to define $c = -\hat{\beta}_1$. One could write a payoff matrix for this model as follows:

$$\begin{bmatrix} -c & -c \\ 0 & 0 \end{bmatrix}$$

The payoff to an individual is $\beta_1 = -c$ with a *p*-score of 1, and 0 with a *p*-score of 0, and the fitness would then be the baseline fitness β_0 plus the payoff.

Model 2, Price equation 2, rule 2. The second is also a linear model, but one in which the cooperation level of the individual itself not only has an effect on her own fitness (typically thought of as negative), but also on the fitness of the relative (typically thought of as positive):

$$w_i = \gamma_{0,0} + \gamma_{1,0} p_i + \gamma_{0,1} q_i + \varepsilon_i$$

In an infinite population, the Generalized Price equation for the second model is

$$\bar{w}\Delta\bar{p} = \hat{\gamma}_{1,0}\mathrm{Var}(p) + \hat{\gamma}_{0,1}\mathrm{Cov}(p,q)$$

where $\hat{\gamma}_{1,0}$ and $\hat{\gamma}_{0,1}$ are independent of the composition of the parent population (represented by p and q), and are equal to $\gamma_{1,0}$ and $\gamma_{0,1}$, respectively. This can also be rewritten as

$$\bar{w}\Delta\bar{p} = \left(\hat{\gamma}_{1,0} + \frac{\mathrm{Cov}(p,q)}{\mathrm{Var}(p)}\hat{\gamma}_{0,1}\right)\mathrm{Var}(p)$$

From this, we can see that $\Delta\bar{p} > 0$ if and only if $\hat{\gamma}_{1,0} + \frac{\mathrm{Cov}(p,q)}{\mathrm{Var}(p)}\hat{\gamma}_{0,1} > 0$. In infinite population models, $\frac{\mathrm{Cov}(p,q)}{\mathrm{Var}(p)}$ will coincide with relatedness r between the individuals that interact, and if we then define $c = -\gamma_{1,0}$ and $b = \gamma_{0,1}$, we can rewrite this as

$$\bar{w}\Delta\bar{p} = (-c + rb)\mathrm{Var}(p)$$

Here we naturally recognize Hamilton's rule. The payoff matrix for model 2 moreover makes this a prisoner's dilemma with equal gains from switching:

$$\begin{bmatrix} b - c & -c \\ b & 0 \end{bmatrix}$$

Model 2, Price equation 1. It is worth noticing that we can also apply the first Price-like equation to the second model. If we do, we get

$$\bar{w}\Delta\bar{p} = \hat{\beta}_1\mathrm{Var}(p)$$

Calculations in Appendix A.2 show that $\hat{\beta}_0 = \gamma_{0,0} + \left(1 - \frac{\mathrm{Cov}(p,q)}{\mathrm{Var}(p)}\right)\bar{p}\gamma_{0,1}$ and $\hat{\beta}_1 = \gamma_{1,0} + \frac{\mathrm{Cov}(p,q)}{\mathrm{Var}(p)}\gamma_{0,1}$ minimize the sum of squared errors for Price-like equation 1 in combination with

Model 2.[1] The first, $\hat{\beta}_0$, does not feature in the Price equation, but $\hat{\beta}_1$ does. Here, I write $\hat{\beta}_1$, because it minimizes the least squares of the errors relative to Model 1. That means it would be equal to β_1, when applied to Model 1, but now that it is applied to Model 2, it does not coincide with any parameter. This Price-like equation is however still an identity, also when applied to the second model, and it still gets the direction of selection right; $\Delta\bar{p} > 0$ if and only if $\hat{\beta}_1 > 0$ (in fact, it would get the direction of selection right for any model that has a constant and a linear term).

The natural interpretation of this rule would be that the effect of having the gene on the individual herself is $\hat{\beta}_1$, which is equal to $\gamma_{1,0} + r\gamma_{0,1}$. This is clearly not the case, unless $\gamma_{0,1} = 0$, in which case we are back in the situation of Model 1. It is however important to note that we now have *two* rules, both of which we applied to Model 2, and both of which get the direction of selection right. There is a silly one, in which we would choose $c = -(\gamma_{1,0} + r\gamma_{0,1})$, and, if one would insist on making it look similar to Hamilton's rule, $b = 0$; and a more sensible one, with $c = -\gamma_{1,0}$ and $b = \gamma_{0,1}$, that everyone would agree is the right one for this model. We will get back to the reasons why we choose Price-like equation 2 here, and not Price-like equation 1, when we compare Models 2 and 3.

Symptoms of underspecification

If we combine the Generalized Price equation for Model 1 with Model 1, in an infinite population, then the sum of squared errors is equal to the variance of the noise term $Var(\varepsilon)$. If we combine the Generalized Price equation for Model 1 with Model 2, in an infinite population, then the sum of squared errors becomes larger than the variance of the noise term $Var(\varepsilon)$. This is the result of under-specification.

In Fig. 4, we see a very simple example of how underspecification can increase the sum of squared errors over the unavoidable level induced by the noise. The example does not illustrate the specific form of underspecification that we get if we leave out the *q*-score as an explanatory variable. Instead, it illustrates the principle by considering an even simpler form of underspecification, in which the *p*-score is left out as an explanatory variable. Both symptoms of underspecification show up here; the sum of squared errors is higher than is justified by the error term of the true model; and the estimator of the coefficient depends on the composition of the parent population. Both of these disappear when we use the correctly specified model instead.

[1] One can also derive $\hat{\beta}_1$ in a more direct way than is done in Appendix A.2. Given that both Price-like equation 1 and Price-like equation 2 are identities, and both have $\bar{w}\Delta\bar{p}$ on the left-hand side, it must be that their right-hand sides are also equal. This implies that $\hat{\beta}_1 \text{Var}(p) = \left(\gamma_{1,0} + \frac{\text{Cov}(p,q)}{\text{Var}(p)}\gamma_{0,1}\right) \text{Var}(p)$, and therefore that $\hat{\beta}_1 = \left(\gamma_{1,0} + \frac{\text{Cov}(p,q)}{\text{Var}(p)}\gamma_{0,1}\right)$. This direct way is possible here because there is only one variable, $\hat{\beta}_1$, and one equation. For some later combinations of models and Price-like equations, there is still one equation, but more degrees of freedom.

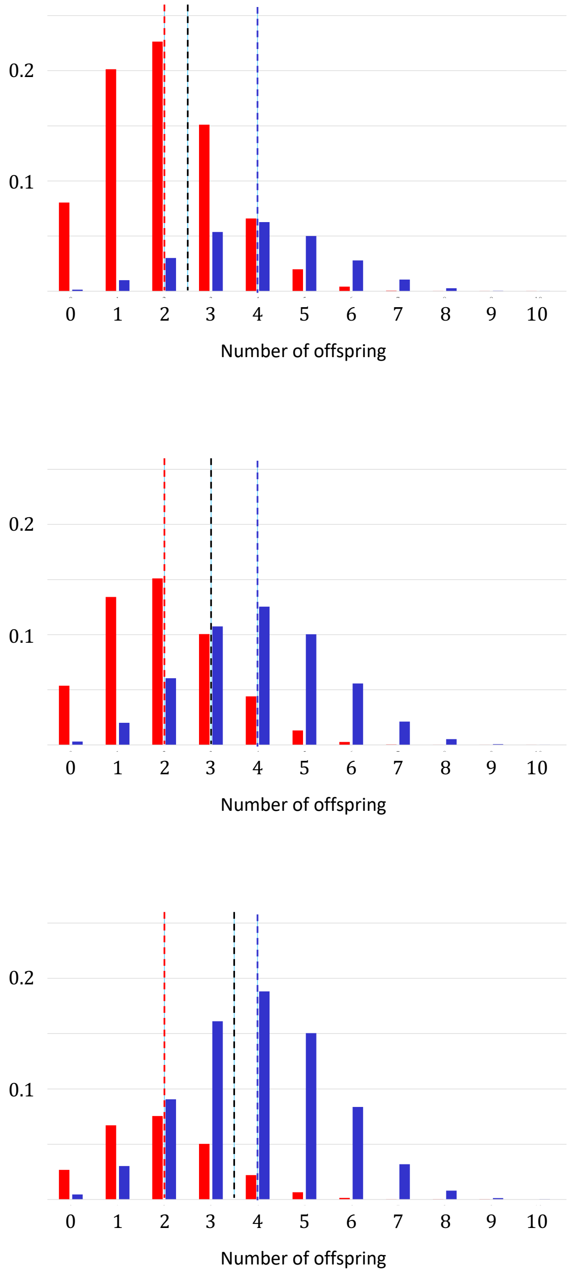

Fig. 4 | Symptoms of underspecification. The red bars indicate frequencies of parents with a *p*-score of 0, and with 0, 1, …, 10 offspring for the example described in the text. The blue bars indicate frequencies of parents with a *p*-score of 1, and with 0, 1, …, 10 offspring. The composition of the parent population differs between the panels; in the top panel, 1 out of 4 parents has a *p*-score of 1; in the middle panel, this is 2 out of 4 parents; and in the bottom panel, it is 3 out of 4 parents. For the correctly specified model $w_i = \alpha + \beta p_i + \varepsilon_i$, the red and the blue dotted lines indicate the values for α and for $\alpha + \beta$, if those are chosen so as to minimize the sum of squared errors relative to the correct model. The α is the expected value of the number of offspring for parents with a *p*-score of 0, while $\alpha + \beta$ is the expected value of the number of offspring for parents with a *p*-score of and 1. The α and β that minimize the sum of squared errors do not depend on the composition of the population. For the mis-specified model $w_i = \alpha + \varepsilon_i$, the black dotted lines indicate the α's for which the sum of squared errors is minimized. These do move around. The sum of squared errors when using the mis-specified model is also much larger at the minimum.

In the example, parents with a *p*-score of 0 draw their offspring from a binomial distribution with 10 trials and a success probability of $\frac{1}{5}$, while parents with a *p*-score of 1 draw their offspring from a binomial distribution, also with 10 trials, but with a success probability of $\frac{2}{5}$. Their *p*-score therefore determines the expected value of their number of offspring; the expected number of offspring is $10 \cdot \frac{1}{5} = 2$ for parents with a *p*-score of 0, and $10 \cdot \frac{2}{5} = 4$ for parents with a *p*-score of 1. If we ignore this, and mis-specify the model by leaving out the *p*-score as an explanatory variable, then the model becomes $w_i = \alpha + \varepsilon_i$. Which value of the α minimizes the sum of squared errors will now depend on the composition of the population. In the first panel, 1 out of 4 parents has a *p*-score of 1, and the sum of squared errors is minimized at $\alpha = 2\frac{1}{2}$; in the second panel, 1 out of 2 parents has a *p*-score of 1, and the sum of squared errors is minimized at $\alpha = 3$; and in the third panel, 3 out of 4 parents has a *p*-score of 1, and the sum of squared errors is minimized at $\alpha = 3\frac{1}{2}$. These are the black dotted lines.

The sum of squared errors is significantly reduced if we do include the *p*-score as an explanatory variable, so the model becomes $w_i = \alpha + \beta p_i + \varepsilon_i$. In that case the expected number of offspring of parents with a *p*-score of 0 is 2, and the expected number of offspring of parents with a *p*-score of 1 is 4, both regardless of the composition of the parent population.

Model 3, Price equation 3, rule 3. The third model allows for an interaction effect between the *p*-score and the *q*-score. This could for instance reflect an efficiency gain from joint cooperation, if positive:

$$w_i = \delta_{0,0} + \delta_{1,0} p_i + \delta_{0,1} q_i + \delta_{1,1} p_i q_i + \varepsilon_i$$

In an infinite population, the Generalized Price equation for the third model is

$$\bar{w}\Delta\bar{p} = \hat{\delta}_{1,0}\mathrm{Var}(p) + \hat{\delta}_{0,1}\mathrm{Cov}(p,q) + \hat{\delta}_{1,1}\mathrm{Cov}(p,pq)$$

where $\hat{\delta}_{1,0}$, $\hat{\delta}_{0,1}$, and $\hat{\delta}_{1,1}$ are independent of the composition of the parent population, represented by p and q, and they are equal to $\delta_{1,0}$, $\delta_{0,1}$, and $\delta_{1,1}$, respectively. This can be rewritten as

$$\bar{w}\Delta\bar{p} = \left(\hat{\delta}_{1,0} + \frac{\mathrm{Cov}(p,q)}{\mathrm{Var}(p)}\hat{\delta}_{0,1} + \frac{\mathrm{Cov}(p,pq)}{\mathrm{Var}(p)}\hat{\delta}_{1,1}\right)\mathrm{Var}(p)$$

From this, we can see that $\Delta\bar{p} > 0$ if and only if $\hat{\delta}_{1,0} + \frac{\mathrm{Cov}(p,q)}{\mathrm{Var}(p)}\hat{\delta}_{0,1} + \frac{\mathrm{Cov}(p,pq)}{\mathrm{Var}(p)}\hat{\delta}_{1,1} > 0$. If we then define $c = -\delta_{1,0}$, $r_{0,1} = \frac{\mathrm{Cov}(p,q)}{\mathrm{Var}(p)}$, $b_{0,1} = \delta_{0,1}$, $r_{1,1} = \frac{\mathrm{Cov}(p,pq)}{\mathrm{Var}(p)}$, and $b_{1,1} = \delta_{1,1}$, then we can rewrite this as

$$\bar{w}\Delta\bar{p} = \left(-c + r_{0,1}b_{0,1} + r_{1,1}b_{1,1}\right)\mathrm{Var}(p)$$

This does not give us the Hamilton's rule we are familiar with, but it does give us a correct criterion for when higher *p*-scores are selected for, in the same way that the derivation of the familiar Hamilton's rule did for Model 2, and the derivation of the first rule did for Model 1; $\Delta\bar{p} > 0$ if and only if $r_{0,1}b_{0,1} + r_{1,1}b_{1,1} > c$. This is the rule suggested by Queller [39], if we assume that genotype and phenotype correlate perfectly.

The payoff matrix for this model is a prisoner's dilemma without equal from switching:

$$\begin{bmatrix} b_{0,1} + b_{1,1} - c & -c \\ b_{0,1} & 0 \end{bmatrix}$$

As before, it is worth noticing that we can apply the earlier Price-like equations to the third model too.

Model 3, Price equation 1. If we apply the Price-like equation that goes with the first model to the third model, we get

$$\bar{w}\Delta\bar{p} = \hat{\beta}_1 \mathrm{Var}(p)$$

with $\hat{\beta}_1 = \delta_{1,0} + \frac{\mathrm{Cov}(p,q)}{\mathrm{Var}(p)}\delta_{0,1} + \frac{\mathrm{Cov}(p,pq)}{\mathrm{Var}(p)}\delta_{1,1}$. Here we use the more direct calculation, suggested in footnote 8 for Price equation 1 applied to Model 2. This Price-like equation is still an identity, also when applied to the third model, and it still gets the direction of selection right; $\Delta\bar{p} > 0$ if and only if $\hat{\beta}_1 > 0$.

The natural interpretation of this rule would be that the effect of having the gene on the individual herself is $\hat{\beta}_1$, which is equal to $\delta_{1,0} + \frac{\mathrm{Cov}(p,q)}{\mathrm{Var}(p)}\delta_{0,1} + \frac{\mathrm{Cov}(p,pq)}{\mathrm{Var}(p)}\delta_{1,1}$ here. This is clearly not the case, unless $\delta_{0,1} = \delta_{1,1} = 0$, in which case we are back in Model 1.

Model 3 with Price equation 2. If we use the Price-like equation that goes with the second model, and apply it to the third model, we get

$$\bar{w}\Delta\bar{p} = \left(\hat{\gamma}_{1,0} + \frac{\mathrm{Cov}(p,q)}{\mathrm{Var}(p)}\hat{\gamma}_{0,1}\right)\mathrm{Var}(p)$$

with

$$\hat{\gamma}_{1,0} = \delta_{1,0} - \delta_{1,1}\left(\frac{\mathrm{Cov}(p,pq)\mathrm{Var}(q) - \mathrm{Cov}(p,q)\mathrm{Cov}(q,pq)}{\left(\mathrm{Cov}(p,q)\right)^2 - \mathrm{Var}(p)\mathrm{Var}(q)}\right)$$

$$\hat{\gamma}_{0,1} = \delta_{0,1} + \delta_{1,1}\left(\frac{\mathrm{Cov}(p,q)\mathrm{Cov}(p,pq) - \mathrm{Cov}(q,pq)\mathrm{Var}(p)}{\left(\mathrm{Cov}(p,q)\right)^2 - \mathrm{Var}(p)\mathrm{Var}(q)}\right)$$

The derivation is a bit long and boring and can be found in Appendix A.3. This Price-like equation is still an identity, also when applied to the third model, and it still gets the direction of selection right; $\Delta\bar{p} > 0$ if and only if $\hat{\gamma}_{1,0} + \frac{\mathrm{Cov}(p,q)}{\mathrm{Var}(p)}\hat{\gamma}_{0,1} > 0$.

The natural interpretation of this rule would be that the effect of having the gene on the individual herself is $\hat{\gamma}_{1,0}$, of which the first formula above describes how it depends on $\delta_{0,1}$ and $\delta_{1,1}$, while the effect of the q-score is $\hat{\gamma}_{0,1}$, of which the first formula above describes how it depends on $\delta_{0,1}$ and $\delta_{1,1}$. This is clearly not the case, unless $\delta_{1,1} = 0$, in which case we are back in the situation of Model 2.

The last expression looks like the traditional Hamilton's rule, with $c = -\hat{\gamma}_{1,0}$ and $b = \hat{\gamma}_{0,1}$. It should be noted, though, that the expressions for the b and the c are not constants, and while they depend on both model parameters, which could in principle be fine, they also depend on population state properties, which is not.

For model 2, we chose rule 2 (the traditional Hamilton's rule), and not rule 1 (the standard rule for non-social traits). The reason to do this, is not that rule 1 gets the direction of selection wrong; it does not. The reason is that the latter mistakenly suggests that having the trait comes with a fitness benefit to oneself, which is not true; it comes with a cost to oneself, and a benefit to the other. The reason it can nonetheless go up in frequency, is that because of relatedness, those that bear the costs are also disproportionately often on the receiving end. Rule 2 does, and rule 1 does not, reflect that. A symptom of the misspecification, if we nonetheless use rule 1, is that $\hat{\beta}_1$ (the estimator of the effect on oneself) depends on the composition of the population.

If we follow the same logic regarding choosing rules for model 3, then we should go with rule 3, that is, $r_{0,1}b_{0,1} + r_{1,1}b_{1,1} > c$, or Queller's rule, and not rule 2, which is the traditional Hamilton's rule. The latter would still get the direction of selection right, but it would mistakenly suggest that the effect of having the trait on the other is independent of whether or not the other has it too (or in other words: it would be blind to the fact that there is an interaction term). A symptom of this misspecification is that $\hat{\gamma}_{1,0}$ and $\hat{\gamma}_{0,1}$ (the estimator of the effect on oneself and the other, respectively) both depend on the composition of the population.

The relation between these examples can be summarized in a table that, for three models and three rules, represents all nine combinations of them.

For the combinations in yellow, the rules are more general than the models need them to be. We did not discuss those combinations above, but they are relatively easy to check. Since the rules are nested, in the sense that rule 1 is a special case of rule 2, and rule 2 is a special case of rule 3, all we need to do is choose zeros for the unused variables. In that sense these combinations do not really produce rules that are different, or invite different interpretations; they are only written a bit more inefficiently, by adding terms that end up being 0. All of these are stable across population states.

That is not true for the combinations in red. There the c and the b's vary with the population state, and the rules are not general enough for their model. For the combinations in green, none of the b's and c's in the rule is 0 (which implies that the rule is not overspecified for the model) and none of them depend on population properties (which implies that they are also not underspecified).

		Model 1:	Model 2:	Model 3:
		$\alpha + \beta p_i$	$\alpha + \gamma_{1,0} p_i + \gamma_{0,1} q_i$	$\alpha + \delta_{1,0} p_i + \delta_{0,1} q_i + \delta_{1,1} p_i q_i$
Rule 1:				
$0 > c$	$c =$	🟩 $-\beta$	🟥 $-\left(\gamma_{1,0} + r_{0,1}\gamma_{0,1}\right)$	🟥 $-\left(\delta_{1,0} + r_{0,1}\delta_{0,1} + r_{1,1}\delta_{1,1}\right)$
Rule 2:				
$rb > c$	$c =$	🟨 $-\beta$	🟩 $-\gamma_{1,0}$	🟥 $-\delta_{1,0} + s_c \delta_{1,1}$
	$b =$	🟨 0	🟩 $\gamma_{0,1}$	🟥 $\delta_{0,1} + s_b \delta_{1,1}$
Rule 3:				
$r_{0,1} b_{0,1} + r_{1,1} b_{1,1} > c$	$c =$	🟨 $-\beta$	🟨 $-\gamma_{1,0}$	🟩 $-\delta_{1,0}$
	$b_{0,1} =$	🟨 0	🟨 $\gamma_{0,1}$	🟩 $\delta_{0,1}$
	$b_{1,1} =$	🟨 0	🟨 0	🟩 $\delta_{1,1}$

Fig. 5|Three rules and three models. All combinations of the three rules and the three models. All rules indicate the direction of selection correctly for all models. Yellow indicates a combination of a rule and a model, where the rule is more general than is needed for the model. This leads to one or more b's being 0. Red indicates a combination of a rule and a model, where the rule is not general enough for the model. This leads to one or more b's and c's that depend on the population state. Terms that depend on the population state are abbreviated as follows: $r = r_{0,1} = \frac{\mathrm{Cov}(p,q)}{\mathrm{Var}(p)}$, $r_{1,1} = \frac{\mathrm{Cov}(p,pq)}{\mathrm{Var}(p)}$, $s_b = \frac{\mathrm{Cov}(p,q)\mathrm{Cov}(p,pq) - \mathrm{Cov}(q,pq)\mathrm{Var}(p)}{\left(\mathrm{Cov}(p,q)\right)^2 - \mathrm{Var}(p)\mathrm{Var}(q)}$, and $s_c = \frac{\mathrm{Cov}(p,pq)\mathrm{Var}(q) - \mathrm{Cov}(p,q)\mathrm{Cov}(q,pq)}{\left(\mathrm{Cov}(p,q)\right)^2 - \mathrm{Var}(p)\mathrm{Var}(q)}$. Rule 1 is the standard rule for non-social traits. Rule 2 is the classical Hamilton's rule. Rule 3 is a rule that allows for an interaction effect. An appropriate name for this rule would be Queller's rule [39]. This rule can in turn be nested in a sequence of ever more general rules if we allow for *p*- and *q*-scores that are not restricted to be binary.

This matrix of combinations of models and rules also helps understand what causes the contentiousness of the debate on the generality of Hamilton's rule, and why there are no signs of convergence. Rule 2, which is the rule that we get from applying the standard Price equation (also known as using the regression method), is a completely general rule, in the sense that whatever the true model is, it always gets the direction of selection right. This is true, but that fact is not a good argument for singling this rule out as more helpful, meaningful, or insightful than other rules. In the example above, we have seen that if we apply the Generalized Price equation (which is also using the regression method, but now with a richer menu of alternative underlying statistical models), then this can also give us Rule 1 or Rule 3, depending on the statistical model we use. These rules are equally correct, in the sense that they also always get the direction of selection right, and they are also equally general, in the sense that they get it right for every possible model. Being a general rule, that always gets the direction of selection right, therefore, cannot be a criterion for elevating Rule 2 (which is the classical Hamilton's rule), above the other ones, because Rules 1 and 3 are also general rules, that always get the direction of selection right.

Since being completely general and always getting the direction of selection right does not single out any of the possible rules, we need additional criteria. A natural criterion would be that besides being correct, the terms in the rule would have to be meaningful. More precisely, we think the rule should do what Rule 1 does for Model 1, and what Rule 2 does for Model 2, and that is to separate model parameters from properties of the population state. As described above, we do not apply Rule 1 – the rule for non-social traits – to a social trait with fitnesses effects described by Model 2, and the reason why we do not do this is not that it does not get the prediction right. It does. The reason why we do not apply Rule 1 to Model 2 is that Rule 1 is mis-specified for Model 2, as it does not reflect the fact that Model 2 allows for a trait with a fitness cost to the individual itself, that is nonetheless selected, because of the positive fitness effects between related individuals that outweigh the negative fitness effect on oneself. In other words, it would be wrong to interpret a negative c in Rule 1, applied to Model 2, as the fitness effect on oneself.

The exact same reason would suggest not to apply Rule 2, which is the traditional Hamilton's rule, to models of social interactions that do not fit Model 2, such as for instance Model 3. The argument against interpreting the c in Rule 1, applied to Model 2, as the fitness effect on oneself carries over to this combination of rule and model, as it implies that we can also not interpret the c and b in Rule 2, when applied to Model 3, as the costs and benefits of the social behaviour.

One of the reasons why the debate in the literature on the generality of Hamilton's rule is so long-lasting, is that it focuses on whether or not rules are correct, and not on whether they are (also) meaningful. One side of the debate has always returned to the argument that Rule 2 is general and correct – whatever the model or the data (see for instance [16], [18], [19], [30], [40]). This is true, but it is not an argument for elevating Rule 2 above other rules. The other side of the debate keeps bringing up models that do not fit Model 2 (see for instance [39], [41], [42]). Sometimes arguments on this side take a completely different approach, and rather than describing selection with the Price equation, and then worry about whether or not one can interpret regression coefficients in it as benefits and costs, they start with models with a priori interpretable definitions for b and c, for instance by using what is called the counterfactual approach, rather than the regression approach, to define costs and benefits ([22], [24]). Starting at the opposite end, this can then result in rules that end up getting the direction of selection wrong.

When Queller's rule [39] was published, this opened the door to the idea that Hamilton's rule was nested in a more general rule, or a more general set of rules. A News and Views responding to it [40] however immediately closed that door again, by claiming that "*the third, synergistic, term in Queller's form can be made to disappear by agreeing to define benefit and cost as the average effects on individual's fitnesses, rather than as arbitrary terms in a model of fitness. So in Queller's simple model, Hamilton's rule, with costs and benefits correctly understood, is perfectly adequate for deciding the direction of change in gene frequency.*"

This is first of all a representative example where Hamilton's rule getting the direction of selection right is treated as a decisive argument in its favour, and as an argument against a different rule, even though this other rule also gets the direction of selection right. Regarding the interpretation, Appendix B4.4 moreover shows that this claim is simply incorrect; the benefits and costs that make Hamilton's rule work are the ones that we get by applying the Generalized Price equation for Model 2 to Model 3, and these are ***not*** the average effects on the individual's fitnesses. Also, there is nothing about the terms in Queller's rule [39] that makes them conceptually any different from the terms in Hamilton's rule. Suggesting that Queller's rule [39] includes terms that are arbitrary in ways that terms in Hamilton's rule are not, therefore, is also incorrect. The literature has nonetheless effectively abandoned this path forward immediately after Queller [39] pointed to it.

3. The general version of Hamilton's rule.

If *p*- and *q*-scores can only be 0 or 1, there are only four possibilities combinations of a *p*- and a *q*-score that one can have. If we denote the expected number of offspring of an individual with a *p*-score of p and a *q*-score of q as $E(w^{p,q})$, then there are also only four of those. In Model 3 above, fitness depends on *p*- and a *q*-scores as follows:

$$w_i = \delta_{0,0} + \delta_{1,0}p_i + \delta_{0,1}q_i + \delta_{1,1}p_iq_i + \varepsilon_i$$

That implies that $E(w^{0,0}) = \delta_{0,0}$, $E(w^{1,0}) = \delta_{0,0} + \delta_{1,0}$, $E(w^{0,1}) = \delta_{0,0} + \delta_{0,1}$, and $E(w^{1,1}) = \delta_{0,0} + \delta_{1,0} + \delta_{0,1} + \delta_{1,1}$. With only four values for $E(w^{p,q})$ and four δ's, there is no scope for introducing more coefficients in the model.

If we also allow for *p*- and *q*-scores other than 0 and 1, fitnesses may depend on those in richer ways than the three models above allow for. Below, we will consider a general set of models, which are described by the coefficients that are included. The set of (non-zero) coefficients is denoted by E, and if $(r, s) \in E$, then the model includes the term $\beta_{r,s}{p_i}^r{q_i}^s$. Given a set of coefficients E, the model would be

$$w_i = \sum_{(k,l) \in E} \beta_{k,l}{p_i}^k{q_i}^l + \varepsilon_i$$

where ε_i is a noise term with expected value 0. Whether including a term in the model would be of added value in describing actual fitness effects in a real-life example is of course a matter of statistics. For every behaviour, there is a point where additional coefficients are not going to be statistically significant, even with a lot of data. Depending on the behaviour that the model aims to describe, it is an empirical question for what set E – or, for what model – that happens. For non-social behaviours, Model 1 above could be enough, which would imply that only the coefficients for ${p_i}^0{q_i}^0 = 1$ and ${p_i}^1{q_i}^0 = p_i$ are needed. Alternatively, it could be that the behaviour is non-social, but the fitness is not linear in the *p*-score. This does of course require non-binary *p*-scores, and if we allow for those, then higher order terms could be included. Those would be terms ${p_i}^i{q_i}^0 = {p_i}^i$. If the *q*-score represents the *p*-score of the interaction partner, and the behaviour is non-social, then the *q*-score does not affect fitness, and hence only coefficients where the exponent of q_i is 0 are included.

For social behaviours, Model 2 could describe the fitness accurately, and for those the set of nonzero coefficients would be $E = \{(0,0), (1,0), (0,1)\}$. Model 3 is also an option, and this would amount to choosing $E = \{(0,0), (1,0), (0,1), (1,1)\}$. This setup however also allows for larger sets E that include coefficients that are not included in Model 3.

The Generalized Price equation in regression form can now be stated for any statistical model, specified by the set of coefficients E we include. If we choose a set E, and we choose the coefficients $\hat{\beta}_{r,s}$ so as to minimize the sum of squared errors relative to the model given by $w_i = \sum_{(k,l)\,\in\,E} \beta_{k,l} {p_i}^k {q_i}^l + \varepsilon_i$, then the Generalized Price equation for this choice of E reads

$$\boxed{\bar{w}\Delta\bar{p} = \left(\sum_{(k,l)\in E} \hat{\beta}_{k,l} \frac{\mathrm{Cov}(p, p^k q^l)}{\mathrm{Var}(p)} \right) \mathrm{Var}(p) + E(w\Delta p)}$$

Here it is worth emphasizing that this produces a Price-like equation for every choice of the set of coefficients E. That implies that one can make any combination of a set of coefficients E that is used for writing a Price-like equation, and a set of coefficients E' for the model that one would want to consider. Even if these sets differ, the Price-like equation for E, applied to a model specified by E', remains an identity. If E is a subset of E', then the Price-like equation is underspecified. If E' is a subset of E, then the Price-like equation is overspecified, and the coefficients that are in E, but not in E', will be 0. It is also possible that E is not a subset of E', and E' is also not a subset of E. For such a combination, one could say that the Price-like equation based on the set of coefficients E, and applied to a model specified by E', is both over- and underspecified. If E and E' coincide – that is, if the model used as a basis for the Price-like equation and the model we are studying are the same – then $\hat{\beta}_{k,l} = \beta_{k,l}$ for all coefficients, or, in other words, the coefficients in the Price-like equation coincide with the coefficients of the model we are considering. This generalizes the point made earlier, which is that all of these Price-like equations are identities, but the regression coefficients $\hat{\beta}_{k,l}$ in it only have a meaningful interpretation if the model is correctly specified.

If we define $r_{k,l} = \frac{\mathrm{Cov}(p,p^r q^s)}{\mathrm{Var}(p)}$, then the Generalized Hamilton's rule is that $\Delta\bar{p} > 0$ if

$$\boxed{\sum_{(k,l)\in E} r_{k,l} b_{k,l} > 0}$$

where $b_{k,l} = \beta_{k,l}$, and $r_{k,l} = \frac{\mathrm{Cov}(p,p^k q^l)}{\mathrm{Var}(p)}$, both for all $(k,l) \in E'$.

Because $r_{0,0} = \frac{\mathrm{Cov}(p,p^0 q^0)}{\mathrm{Var}(p)} = \frac{\mathrm{Cov}(p,1)}{\mathrm{Var}(p)} = 0$, it does not matter if we leave the coefficient for $(0,0)$ in or take it out. We can moreover write c for $-b_{1,0} = -\hat{\beta}_{1,0}$, and put it on the right-hand side of this inequality.

The rule for selection of non-social traits with linear effects is a special case, if we choose the model that only contains $\beta_{1,0}$. Because $r_{1,0} = \frac{\mathrm{Cov}(p,p^1 q^0)}{\mathrm{Var}(p)} = \frac{\mathrm{Var}(p)}{\mathrm{Var}(p)} = 1$, this produces the rule that a trait is selected if it increases an individual's fitness, or, in other words, if $\hat{\beta}_{1,0} > 0$.

This still requires that the model is correctly specified, because only then do we have $\hat{\beta}_{1,0} = \beta_{1,0}$.

For non-social traits with (also) quadratic fitness effects, we would have to allow $\beta_{1,0}$ and $\beta_{2,0}$ to be non-zero. The rule then is $b_{1,0} + b_{2,0}\frac{\mathrm{Cov}(p,p^2)}{\mathrm{Var}(p)} > 0$. If we additionally assume random mating, then this simplifies to $b_{1,0} + b_{2,0}\left(\frac{1}{2} + p\right)$, as we have seen in Section 5 of the twin TI discussion paper on the Generalized Price equation.

The original Hamilton's rule is also a special case, if we choose the model that contains only $\beta_{1,0}$ and $\beta_{0,1}$. Because $r_{0,1} = \frac{\mathrm{Cov}(p,p^0q^1)}{\mathrm{Var}(p)} = \frac{\mathrm{Cov}(p,q)}{\mathrm{Var}(p)}$, the rule then becomes that $\Delta\bar{p} > 0$ if $r_{1,0}b_{1,0} + r_{0,1}b_{0,1} > 0$, and with $r_{1,0} = 1$, $c = -\hat{\beta}_{1,0}$, $r = r_{0,1}$, and $b = \hat{\beta}_{0,1}$, this gives us the Hamilton's rule we are familiar with. This also still requires that the model is correctly specified, because only then do we have $\hat{\beta}_{1,0} = \beta_{1,0}$ and $\hat{\beta}_{0,1} = \beta_{0,1}$.

The message from the Generalized Price equation now carries over to the Generalized Hamilton's rule. There is in fact a Hamilton-like rule for every choice of E, and they all get the direction of selection right, but only if the set of coefficients matches the set of coefficients of the model we are considering, do they have a meaningful interpretation.

4. Relatedness

There are two final remarks to be made regarding relatedness.

The first is that in models, both the classic relatedness r from Hamilton's original rule, and the general relatednesses $r_{k,l} = \frac{\mathrm{Cov}(p,p^rq^s)}{\mathrm{Var}(p)}$, of which the classical relatedness is a special case ($r = r_{0,1}$), are properties of the population structure. That means that in models where we assume infinitely large populations, the law of large numbers guarantees that the randomness disappears, and probabilities and frequencies coincide. For instance, if we consider full siblings, and assume random mating, then that implies that the probability that one sibling has a gene if the other has it, is $(1-r)\bar{p} + r = \frac{1}{2}\bar{p} + \frac{1}{2}$, where $\bar{p}$ is the share of the population that has it (see Appendix B4). The model assumptions, including the assumption of an infinitely large population, then imply that the share of individuals whose sibling has the gene, out of those that have it themselves, is in fact $\frac{1}{2}\bar{p} + \frac{1}{2}$.

With data, on the other hand, the $\frac{\mathrm{Cov}(p,p^rq^s)}{\mathrm{Var}(p)}$ term is an *estimator* of a property of the true, underlying population structure. With few observations, this estimator will be off, on average, more than with many observations. In the twin Tinbergen discussion paper on the Generalized Price equation, we stressed that this is true for the coefficients in it (the β's, γ's, and δ's), when considering data, but it is worth stressing that this is also true for the terms $\frac{\mathrm{Cov}(p,p^rq^s)}{\mathrm{Var}(p)}$, when these relate to properties of the population structure.

The second remark is that relatedness $r = r_{0,1}$ from the classical Hamilton's rule is frequency-independent in infinite population models. This is not true for other properties of the population structure, which do determine relevant properties of population states at different frequencies. In Appendix B4.1 we for instance calculate that

$$r_{1,1} = (1 - r)\bar{p} + r$$

which changes with $\bar{p}$. That implies that, while the rule for selection of non-social traits and the classical Hamilton's rule either predict one of the other to go to fixation, richer models allow for richer equilibrium properties, such as coexistence or bistability. For coefficients that are actual constants, that requires the relatedness coefficients to vary with $\bar{p}$, which is not the case for $r_{0,1}$, but it is for the others.

Appendix

A.1 Price-like equation 1, applied to Model 1

As a finger exercise, we first apply Price-like equation 1 to Model 1. That helps make a few observations that are relevant, when we minimize squared errors, not in a statistical setting, but in a setting with a model.

We get Price-like equation 1 from minimizing the sum of squared errors, relative to Model 1. The true parameters are indicated by β_0 and β_1, their estimates by $\hat{\beta}_0$ and $\hat{\beta}_1$. The actual errors are indicated by ε_i, and the errors, the squared sum of which we minimize, are indicated by $\hat{\varepsilon}_i$.

The sum of squared errors is

$$\sum_{i=1}^{n} \hat{\varepsilon}_i{}^2 = \sum_{i=1}^{n} \left(w_i - (\hat{\beta}_0 + \hat{\beta}_1 p_i)\right)^2 = \sum_{i=1}^{n} \left(\beta_0 + \beta_1 p_i + \varepsilon_i - (\hat{\beta}_0 + \hat{\beta}_1 p_i)\right)^2$$

If we minimize this, we begin with setting the derivative to $\hat{\beta}_0$ to 0.

$$-2\sum_{i=1}^{n} \left(\beta_0 + \beta_1 p_i + \varepsilon_i - (\hat{\beta}_0 + \hat{\beta}_1 p_i)\right) = 0$$

$$\sum_{i=1}^{n} (\beta_0 - \hat{\beta}_0) + \sum_{i=1}^{n} (\beta_1 - \hat{\beta}_1) p_i + \sum_{i=1}^{n} \varepsilon_i = 0$$

$$\beta_0 - \hat{\beta}_0 + (\beta_1 - \hat{\beta}_1)\bar{p} + \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i = 0$$

Then we set the derivative to $\hat{\beta}_1$ to 0

$$-2\sum_{i=1}^{n} p_i \left(\beta_0 + \beta_1 p_i + \varepsilon_i - (\hat{\beta}_0 + \hat{\beta}_1 p_i)\right) = 0$$

$$\sum_{i=1}^{n} (\beta_0 - \hat{\beta}_0)\, p_i + \sum_{i=1}^{n} (\beta_1 - \hat{\beta}_1) p_i{}^2 + \sum_{i=1}^{n} \varepsilon_i p_i = 0$$

$$(\beta_0 - \hat{\beta}_0)\bar{p} + (\beta_1 - \hat{\beta}_1)\frac{1}{n}\sum_{i=1}^{n} p_i{}^2 + \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i p_i = 0$$

Although outside the Price equation literature, it is unusual to apply minimization of squared errors in a modeling context – where we know what the true model is, and therefore what

β_0 and β_1 are – we can nonetheless still do it. In an infinite population, we can use that the errors have mean 0, which implies that $\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i = 0$, and we can use that the errors are uncorrelated with the *p*-scores, which implies that $\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\, p_i = 0$. Then we have two equations:

$$\beta_0 - \hat{\beta}_0 + (\beta_1 - \hat{\beta}_1)\bar{p} = 0$$

$$(\beta_0 - \hat{\beta}_0)\bar{p} + (\beta_1 - \hat{\beta}_1)\frac{1}{n}\sum_{i=1}^{n} {p_i}^2 = 0$$

It is clear that both of these hold if $\hat{\beta}_0 = \beta_0$ and $\hat{\beta}_1 = \beta_1$. In a statistical context, one cannot do this, because there we do not know the β_0 and β_1, and the purpose of the estimation procedure is to get an estimate of those. Here, however, we do know what the actual model is, and we can simply pick $\hat{\beta}_0 = \beta_0$ and $\hat{\beta}_1 = \beta_1$.

Of course, on its own, this is a futile exercise, as it unveils something that is not veiled. It does however illustrate that Price equation 1 applied to Model 1 gets the model right. This is an instructive benchmark for when we go on to allow for Price-like equations that are applied to non-matching models. An observation that may also help comparisons with this benchmark, is that the errors that are found by the minimization coincide with the actual errors. Here we use that we chose $\hat{\beta}_0 = \beta_0$ and $\hat{\beta}_1 = \beta_1$.

$$\hat{\varepsilon}_i = w_i - (\hat{\beta}_0 + \hat{\beta}_1 p_i) = \beta_0 + \beta_1 p_i + \varepsilon_i - (\hat{\beta}_0 + \hat{\beta}_1 p_i) = \varepsilon_i$$

for $i = 1, .., n$.

If we would leave out the actual errors ε_i at the very beginning, then the calculations come out the same. This would amount to assuming that $\varepsilon_i = 0$ for all i, which is more than just assuming that errors have mean 0 (which implies that $\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i = 0$) and that they are uncorrelated with *p*-scores (which implies that $\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i p_i = 0$). It can be seen, though, as a shortcut for assuming infinitely large populations, as we do here. In that case, we would get $\widehat{\varepsilon_i} = 0$ for $i = 1, .., n$.

I will not repeat this for Price equation 2 applied to Model 2, or Price equation 3 applied to Model 3, as I hope it is clear that all minimizations will result in the estimators coinciding with the actual coefficients. This is also verified in the shortcuts provided in Appendix B below.

A.2 Price equation 1, applied to Model 2

We now apply Price-like equation 1 to Model 2. The sum of squared errors is

$$\sum_{i=1}^{n} \hat{\varepsilon}_i{}^2 = \sum_{i=1}^{n} \left(w_i - (\hat{\beta}_0 + \hat{\beta}_1 p_i)\right)^2 = \sum_{i=1}^{n} \left(\gamma_{0,0} + \gamma_{1,0} p_i + \gamma_{0,1} q_i + \varepsilon_i - (\hat{\beta}_0 + \hat{\beta}_1 p_i)\right)^2$$

If we minimize this, we begin with setting the derivative to $\hat{\beta}_0$ to 0.

$$-2 \sum_{i=1}^{n} \left(\gamma_{0,0} + \gamma_{1,0} p_i + \gamma_{0,1} q_i + \varepsilon_i - (\hat{\beta}_0 + \hat{\beta}_1 p_i)\right) = 0$$

$$n(\gamma_{0,0} - \hat{\beta}_0) + (\gamma_{1,0} - \hat{\beta}_1) \sum_{i=1}^{n} p_i + \gamma_{0,1} \sum_{i=1}^{n} q_i + \sum_{i=1}^{n} \varepsilon_i = 0$$

$$(\gamma_{0,0} - \hat{\beta}_0) + (\gamma_{1,0} + \gamma_{0,1} - \hat{\beta}_1)\bar{p} + \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i = 0$$

In the last step, we use that $\bar{p} = \bar{q}$. Then we set the derivative to $\hat{\beta}_1$ to 0

$$-2 \sum_{i=1}^{n} p_i \left(\gamma_{0,0} + \gamma_{1,0} p_i + \gamma_{0,1} q_i + \varepsilon_i - (\hat{\beta}_0 + \hat{\beta}_1 p_i)\right) = 0$$

$$(\gamma_{0,0} - \hat{\beta}_0) \sum_{i=1}^{n} p_i + (\gamma_{1,0} - \hat{\beta}_1) \sum_{i=1}^{n} p_i{}^2 + \gamma_{0,1} \sum_{i=1}^{n} p_i q_i + \sum_{i=1}^{n} \varepsilon_i p_i = 0$$

$$(\gamma_{0,0} - \hat{\beta}_0)\bar{p} + (\gamma_{1,0} - \hat{\beta}_1) \frac{1}{n} \sum_{i=1}^{n} p_i{}^2 + \gamma_{0,1} \frac{1}{n} \sum_{i=1}^{n} p_i q_i + \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i p_i = 0$$

In an infinite population, we can use that the errors are assumed to have mean 0, which implies that $\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i$ becomes 0 for large n. We also use that the errors are uncorrelated with the *p*-scores, which implies that $\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i p_i$ also becomes 0 for large n. The we multiply the first equation by $\bar{p}$, which gives us two equations:

$$(\gamma_{0,0} - \hat{\beta}_0)\bar{p} + (\gamma_{1,0} + \gamma_{0,1} - \hat{\beta}_1)\bar{p}^2 = 0$$

and

$$(\gamma_{0,0} - \hat{\beta}_0)\bar{p} + (\gamma_{1,0} - \hat{\beta}_1) \frac{1}{n} \sum_{i=1}^{n} p_i{}^2 + \gamma_{0,1} \frac{1}{n} \sum_{i=1}^{n} p_i q_i = 0$$

These combine to

$$(\gamma_{1,0}+\gamma_{0,1}-\hat{\beta}_1)\bar{p}^2 = (\gamma_{1,0}-\hat{\beta}_1)\frac{1}{n}\sum_{i=1}^{n} {p_i}^2 + \gamma_{0,1}\frac{1}{n}\sum_{i=1}^{n} p_i q_i$$

$$0 = (\gamma_{1,0}-\hat{\beta}_1)\left(\frac{1}{n}\sum_{i=1}^{n} {p_i}^2 - \frac{1}{n^2}\left(\sum_{i=1}^{n} p_i\right)^2\right) + \gamma_{0,1}\left(\frac{1}{n}\sum_{i=1}^{n} p_i q_i - \frac{1}{n^2}\left(\sum_{i=1}^{n} p_i\right)^2\right)$$

With $\sum_{i=1}^{n} p_i = \sum_{i=1}^{n} q_i$, this is also

$$0 = (\gamma_{1,0}-\beta_1)\mathrm{Var}(p) + \gamma_{0,1}\mathrm{Cov}(p,q)$$

or

$$\boxed{\hat{\beta}_1 = \gamma_{1,0} + \gamma_{0,1}\frac{\mathrm{Cov}(p,q)}{\mathrm{Var}(p)} = \gamma_{1,0} + r\gamma_{0,1}}$$

Then combining with the first equation gives

$$(\gamma_{0,0}-\hat{\beta}_0) + \left(\gamma_{1,0}+\gamma_{0,1}-(\gamma_{1,0}+\gamma_{0,1}r)\right)\bar{p} = 0$$

$$\boxed{\hat{\beta}_0 = \gamma_{0,0} + (1-r)\bar{p}\gamma_{0,1}}$$

If we now go back to the errors, we find that

$$\begin{aligned}\hat{\varepsilon}_i &= w_i - (\hat{\beta}_0 + \hat{\beta}_1 p_i)\\ &= \gamma_{0,0} + \gamma_{1,0}p_i + \gamma_{0,1}q_i + \varepsilon_i - \left(\gamma_{0,0} + (1-r)\bar{p}\gamma_{0,1} + (\gamma_{1,0} + r\gamma_{0,1})p_i\right)\\ &= \gamma_{0,1}\left(q_i - \left((1-r)\bar{p} + rp_i\right)\right) + \varepsilon_i\end{aligned}$$

for all $i = 1,..,n$. The errors, the squared sum of which is minimized, therefore do not coincide with the actual errors, and include $\gamma_{0,1}$ times the gap between q_i and $(1-r)\bar{p} + rp_i$. The first is the *q*-score of individual i, the second would be a predictor for q_i based on $\bar{p}$ and p_i. This part represents the misspecification-induced part of the estimated errors.

If we again assume $\varepsilon_i = 0$ for all i as a shortcut for assuming an infinitely large population, then $\hat{\varepsilon}_i = \gamma_{0,1}\left(q_i - \left((1-r)\bar{p} + rp_i\right)\right)$ only reflects the misspecification.

A.3 Price equation 2, applied to Model 3

We now apply Price-like equation 2 to Model 3. The derivation is rather boring, but there is no way around it (see Footnote 8). The sum of squared errors is

$$\sum_{i=1}^{n} \hat{\varepsilon}_i{}^2 = \sum_{i=1}^{n} \left(\delta_{0,0} + \delta_{1,0}p_i + \delta_{0,1}q_i + \delta_{1,1}p_iq_i + \varepsilon_i - \left(\hat{\gamma}_{0,0} + \hat{\gamma}_{1,0}p_i + \hat{\gamma}_{0,1}q_i\right)\right)^2$$

If we minimize this, we begin with setting the derivative to $\hat{\gamma}_{0,0}$ to 0.

$$-2\sum_{i=1}^{n} \left(\delta_{0,0} + \delta_{1,0}p_i + \delta_{0,1}q_i + \delta_{1,1}p_iq_i + \varepsilon_i - \left(\hat{\gamma}_{0,0} + \hat{\gamma}_{1,0}p_i + \hat{\gamma}_{0,1}q_i\right)\right) = 0$$

$$n\left(\delta_{0,0} - \hat{\gamma}_{0,0}\right) + \left(\delta_{1,0} - \hat{\gamma}_{1,0}\right)\sum_{i=1}^{n} p_i + \left(\delta_{0,1} - \hat{\gamma}_{0,1}\right)\sum_{i=1}^{n} q_i + \delta_{1,1}\sum_{i=1}^{n} p_iq_i + \sum_{i=1}^{n} \varepsilon_i = 0$$

Then we set the derivative to $\hat{\gamma}_{1,0}$ to 0

$$-2\sum_{i=1}^{n} p_i \left(\delta_{0,0} + \delta_{1,0}p_i + \delta_{0,1}q_i + \delta_{1,1}p_iq_i + \varepsilon_i - \left(\hat{\gamma}_{0,0} + \hat{\gamma}_{1,0}p_i + \hat{\gamma}_{0,1}q_i\right)\right) = 0$$

$$\left(\delta_{0,0} - \hat{\gamma}_{0,0}\right)\sum_{i=1}^{n} p_i + \left(\delta_{1,0} - \hat{\gamma}_{1,0}\right)\sum_{i=1}^{n} p_i{}^2 + \left(\delta_{0,1} - \hat{\gamma}_{0,1}\right)\sum_{i=1}^{n} p_iq_i + \delta_{1,1}\sum_{i=1}^{n} p_i{}^2q_i + \sum_{i=1}^{n} \varepsilon_ip_i = 0$$

Finally we set the derivative to $\hat{\gamma}_{0,1}$ to 0

$$-2\sum_{i=1}^{n} q_i \left(\delta_{0,0} + \delta_{1,0}p_i + \delta_{0,1}q_i + \delta_{1,1}p_iq_i + \varepsilon_i - \left(\gamma_{0,0} + \hat{\gamma}_{1,0}p_i + \hat{\gamma}_{0,1}q_i\right)\right) = 0$$

$$\left(\delta_{0,0} - \hat{\gamma}_{0,0}\right)\sum_{i=1}^{n} q_i + \left(\delta_{1,0} - \hat{\gamma}_{1,0}\right)\sum_{i=1}^{n} p_iq_i + \left(\delta_{0,1} - \hat{\gamma}_{0,1}\right)\sum_{i=1}^{n} q_i{}^2 + \delta_{1,1}\sum_{i=1}^{n} p_iq_i{}^2 + \sum_{i=1}^{n} \varepsilon_iq_i = 0$$

If we then use that $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_ip_i = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_iq_i = 0$, we can rewrite these three equations as

$$\left(\delta_{0,0} - \hat{\gamma}_{0,0}\right) + \left(\delta_{1,0} - \hat{\gamma}_{1,0}\right)\overline{p} + \left(\delta_{0,1} - \hat{\gamma}_{0,1}\right)\overline{q} + \delta_{1,1}\overline{pq} = 0$$

$$\left(\delta_{0,0} - \hat{\gamma}_{0,0}\right)\overline{p} + \left(\delta_{1,0} - \hat{\gamma}_{1,0}\right)\overline{p^2} + \left(\delta_{0,1} - \hat{\gamma}_{0,1}\right)\overline{pq} + \delta_{1,1}\overline{p^2q} = 0$$

$$\left(\delta_{0,0} - \hat{\gamma}_{0,0}\right)\overline{q} + \left(\delta_{1,0} - \hat{\gamma}_{1,0}\right)\overline{pq} + \left(\delta_{0,1} - \hat{\gamma}_{0,1}\right)\overline{q^2} + \delta_{1,1}\overline{pq^2} = 0$$

If we multiply the first equation by $\bar{p}$, it becomes

$$\left(\delta_{0,0}-\hat{\gamma}_{0,0}\right)\bar{p}+\left(\delta_{1,0}-\hat{\gamma}_{1,0}\right)(\bar{p})^2+\left(\delta_{0,1}-\hat{\gamma}_{0,1}\right)(\bar{p})(\bar{q})+\delta_{1,1}(\bar{p})(\overline{pq})=0$$

Combining it with the second equation, we get

$$\left(\delta_{1,0}-\hat{\gamma}_{1,0}\right)\left((\bar{p})^2-\overline{p^2}\right)+\left(\delta_{0,1}-\hat{\gamma}_{0,1}\right)\left((\bar{p})(\bar{q})-\overline{pq}\right)+\delta_{1,1}\left((\bar{p})(\overline{pq})-\overline{p^2q}\right)=0$$

If we multiply the first equation by $\bar{q}$, and combine it with the third equation, we get

$$\left(\delta_{1,0}-\hat{\gamma}_{1,0}\right)\left((\bar{p})(\bar{q})-\overline{pq}\right)+\left(\delta_{0,1}-\hat{\gamma}_{0,1}\right)\left((\bar{q})(\bar{q})-\overline{q^2}\right)+\delta_{1,1}\left((\bar{q})(\overline{pq})-\overline{pq^2}\right)=0$$

This can be written more succinctly as

$$\left(\delta_{1,0}-\hat{\gamma}_{1,0}\right)Var(p)+\left(\delta_{0,1}-\hat{\gamma}_{0,1}\right)\mathrm{Cov}(p,q)+\delta_{1,1}Cov(p,pq)=0$$

$$\left(\delta_{1,0}-\hat{\gamma}_{1,0}\right)Cov(p,q)+\left(\delta_{0,1}-\hat{\gamma}_{0,1}\right)Var(q)+\delta_{1,1}Cov(q,pq)=0$$

If we multiply the second equation with $\frac{\mathrm{Var}(p)}{\mathrm{Cov}(p,q)}$, it becomes

$$\left(\delta_{1,0}-\hat{\gamma}_{1,0}\right)Var(p)+\left(\delta_{0,1}-\hat{\gamma}_{0,1}\right)\frac{\mathrm{Var}(q)\mathrm{Var}(p)}{\mathrm{Cov}(p,q)}+\delta_{1,1}\frac{\mathrm{Cov}(q,pq)\mathrm{Var}(p)}{\mathrm{Cov}(p,q)}=0$$

Combining that with the first equation gives us

$$\begin{aligned}&\left(\delta_{0,1}-\hat{\gamma}_{0,1}\right)\mathrm{Cov}(p,q)+\delta_{1,1}\mathrm{Cov}(p,pq)\\&\qquad=\left(\delta_{0,1}-\hat{\gamma}_{0,1}\right)\frac{\mathrm{Var}(q)\mathrm{Var}(p)}{\mathrm{Cov}(p,q)}+\delta_{1,1}\frac{\mathrm{Cov}(q,pq)\mathrm{Var}(p)}{\mathrm{Cov}(p,q)}\end{aligned}$$

which we rewrite as

$$\left(\delta_{0,1}-\hat{\gamma}_{0,1}\right)\left(\mathrm{Cov}(p,q)-\frac{\mathrm{Var}(q)\mathrm{Var}(p)}{\mathrm{Cov}(p,q)}\right)+\delta_{1,1}\left(\mathrm{Cov}(p,pq)-\frac{\mathrm{Cov}(q,pq)\mathrm{Var}(p)}{\mathrm{Cov}(p,q)}\right)=0$$

$$\left(\delta_{0,1}-\hat{\gamma}_{0,1}\right)+\delta_{1,1}\left(\frac{\mathrm{Cov}(p,pq)-\frac{\mathrm{Cov}(q,pq)\mathrm{Var}(p)}{\mathrm{Cov}(p,q)}}{\mathrm{Cov}(p,q)-\frac{\mathrm{Var}(q)\mathrm{Var}(p)}{\mathrm{Cov}(p,q)}}\right)=0$$

$$\delta_{0,1}+\delta_{1,1}\left(\frac{\mathrm{Cov}(p,pq)-\frac{\mathrm{Cov}(q,pq)\mathrm{Var}(p)}{\mathrm{Cov}(p,q)}}{\mathrm{Cov}(p,q)-\frac{\mathrm{Var}(q)\mathrm{Var}(p)}{\mathrm{Cov}(p,q)}}\right)=\hat{\gamma}_{0,1}$$

$$\boxed{\hat{\gamma}_{0,1}=\delta_{0,1}+\delta_{1,1}\left(\frac{\mathrm{Cov}(p,q)\mathrm{Cov}(p,pq)-\mathrm{Cov}(q,pq)\mathrm{Var}(p)}{\left(\mathrm{Cov}(p,q)\right)^2-\mathrm{Var}(p)\mathrm{Var}(q)}\right)}$$

This means we found $\hat{\gamma}_{0,1}$. In order to also find $\hat{\gamma}_{1,0}$, we fill this in in the equation

$$\left(\delta_{1,0} - \hat{\gamma}_{1,0}\right)\mathrm{Var}(p) + \left(\delta_{0,1} - \hat{\gamma}_{0,1}\right)\mathrm{Cov}(p,q) + \delta_{1,1}\mathrm{Cov}(p,pq) = 0$$

which gives us

$$\left(\delta_{1,0} - \hat{\gamma}_{1,0}\right)\mathrm{Var}(p) - \delta_{1,1}\left(\frac{\mathrm{Cov}(p,q)\mathrm{Cov}(p,pq) - \mathrm{Cov}(q,pq)\mathrm{Var}(p)}{\left(\mathrm{Cov}(p,q)\right)^2 - \mathrm{Var}(p)\mathrm{Var}(q)}\right)\mathrm{Cov}(p,q) \\ + \delta_{1,1}\mathrm{Cov}(p,pq) = 0$$

This we can rewrite as follows

$$\left(\delta_{1,0} - \hat{\gamma}_{1,0}\right)Var(p) - \delta_{1,1}\left(\frac{\left(\mathrm{Cov}(p,q)\right)^2\mathrm{Cov}(p,pq) - \mathrm{Cov}(p,q)\mathrm{Cov}(q,pq)\mathrm{Var}(p)}{\left(\mathrm{Cov}(p,q)\right)^2 - \mathrm{Var}(p)\mathrm{Var}(q)}\right) \\ + \delta_{1,1}\left(\frac{\left(\mathrm{Cov}(p,q)\right)^2\mathrm{Cov}(p,pq) - \mathrm{Cov}(p,pq)\mathrm{Var}(p)\mathrm{Var}(q)}{\left(\mathrm{Cov}(p,q)\right)^2 - \mathrm{Var}(p)\mathrm{Var}(q)}\right) = 0$$

$$\left(\delta_{1,0} - \hat{\gamma}_{1,0}\right)\mathrm{Var}(p) - \delta_{1,1}\left(\frac{\mathrm{Cov}(p,pq)\mathrm{Var}(p)\mathrm{Var}(q) - \mathrm{Cov}(p,q)\mathrm{Cov}(q,pq)\mathrm{Var}(p)}{\left(\mathrm{Cov}(p,q)\right)^2 - \mathrm{Var}(p)\mathrm{Var}(q)}\right) = 0$$

$$\left(\delta_{1,0} - \hat{\gamma}_{1,0}\right) - \delta_{1,1}\left(\frac{\mathrm{Cov}(p,pq)\mathrm{Var}(q) - \mathrm{Cov}(p,q)\mathrm{Cov}(q,pq)}{\left(\mathrm{Cov}(p,q)\right)^2 - \mathrm{Var}(p)\mathrm{Var}(q)}\right) = 0$$

And thereby we found

$$\boxed{\hat{\gamma}_{1,0} = \delta_{1,0} - \delta_{1,1}\left(\frac{\mathrm{Cov}(p,pq)\mathrm{Var}(q) - \mathrm{Cov}(p,q)\mathrm{Cov}(q,pq)}{\left(\mathrm{Cov}(p,q)\right)^2 - \mathrm{Var}(p)\mathrm{Var}(q)}\right)}$$

B. Shortcuts and simple checks

Given that we know that for every choice for a model, the Generalized Price equation has $\bar{w}\Delta\bar{p}$ on the left-hand side, and something times $\mathrm{Var}(p)$ on the right-hand side, we can also, for an infinitely large population, fill in the model to find an expression for $\Delta\bar{p}$, multiply it by $\bar{w}$ and divide by $\mathrm{Var}(p)$ to recover what the core term in the Price-like equation would have to amount to. This is what we will do here, and it works for Price-like equations that are applied to the matching model. We begin with some additional details in describing states of the infinite population for the simple setting with asexual reproduction and a binary *p*-score.

B.1 Population states

With asexual reproduction and a binary *p*-score, a population state in an infinite population model is characterized by the share of individuals that has a *p*-score of 1, which is $\bar{p}$. Every individual is matched with one partner, and since the *q*-score represents the *p*-score of the partner, the first observation is that these averages must be the same; $\bar{p} = \bar{q}$.

The second observation is that, with a binary *p*-score $\mathrm{E}(p^2) = \mathrm{E}(p)$, and hence

$$\mathrm{Var}(p) = \mathrm{E}(p^2) - \mathrm{E}^2(p) = \mathrm{E}(p) - \mathrm{E}^2(p) = \bar{p}(1-\bar{p})$$

Similarly, $\mathrm{Var}(p) = \mathrm{Var}(q)$ if every individual is matched with one partner. We will also refer to those with a *p*-score of 1 as cooperators and those with a *p*-score of 0 as defectors. This does not describe the trait very well in Model 1, but it is consistent across models.

The population structure is characterized by relatedness r, which determines the probabilities with which the two types are matched with themselves and each other.

$$P(q_i = 1 \mid p_i = 1) = (1-r)\bar{p} + r$$

$$P(q_i = 1 \mid p_i = 0) = (1-r)\bar{p}$$

$$P(q_i = 0 \mid p_i = 0) = (1-r)(1-\bar{p}) + r$$

$$P(q_i = 0 \mid p_i = 1) = (1-r)(1-\bar{p})$$

Another way to write this is $r = P(q_i = 1 \mid p_i = 1) - P(q_i = 1 \mid p_i = 0)$, which justifies thinking of r as the additional probability of being matched to a type, if it is the same type as one is oneself (see [24]).

In an infinite population, these probabilities determine the shares of different types of pairs in a straightforward way. If f_0 is the share of pairs in which both have a *p*-score of 0, f_1 is the share of pairs in which one has a *p*-score of 0, and the other a *p*-score of 1, and f_2 is the share of pairs in which both have a *p*-score of 1, then these shares are given by

$$f_0 = (1-r)(1-\bar{p})^2 + r(1-\bar{p})$$

$$f_1 = (1-r)2\bar{p}(1-\bar{p})$$

$$f_2 = (1-r)\bar{p}^2 + r\bar{p}$$

This implies that, in this simple setting, with a binary *p*-score, and every individual being matched with one other individual, there is a straightforward expression for $\mathrm{Cov}(p,q)$.

$$\begin{aligned}\mathrm{Cov}(p,q) = \mathrm{E}(pq) - \mathrm{E}(p)\mathrm{E}(q) = f_2 - \bar{p}\bar{q} = f_2 - \bar{p}^2 = (1-r)\bar{p}^2 + r\bar{p} - \bar{p}^2 = r\bar{p} - r\bar{p}^2 \\ = r\bar{p}(1-\bar{p})\end{aligned}$$

This is useful for calculating r in Model 2, which is also $r_{0,1}$ in Model 3:

$$r = r_{0,1} = \frac{\mathrm{Cov}(p,q)}{\mathrm{Var}(p)} = \frac{r\bar{p}(1-\bar{p})}{\bar{p}(1-\bar{p})}$$

For calculating $r_{1,1}$ in Model 3, and for calculating s_b and s_c, which feature in the condition if we apply Price-like equation 2 to Model 3, we will also need the following covariances. Here we again use the fact that they are binary, which means that, for instance p^2q is 1 if p and q are both 1, and 0 in all other cases.

$$\begin{aligned}\mathrm{Cov}(p,pq) = \mathrm{E}(p^2q) - \mathrm{E}(p)\mathrm{E}(pq) = f_2 - \bar{p}f_2 = (1-\bar{p})f_2 = (1-\bar{p})\big((1-r)\bar{p}^2 + r\bar{p}\big) \\ = \bar{p}(1-\bar{p})\big((1-r)\bar{p} + r\big)\end{aligned}$$

With $\bar{p} = \bar{q}$, also

$$\mathrm{Cov}(q,pq) = \mathrm{E}(pq^2) - \mathrm{E}(q)\mathrm{E}(pq) = f_2 - \bar{q}f_2 = f_2 - \bar{p}f_2 = \mathrm{Cov}(p,pq)$$

These are useful for computing the $r_{1,1}$, which features in Model 3.

$$r_{1,1} = \frac{\mathrm{Cov}(p,pq)}{\mathrm{Var}(p)} = \frac{\bar{p}(1-\bar{p})\big((1-r)\bar{p}+r\big)}{\bar{p}(1-\bar{p})} = \big((1-r)\bar{p}+r\big)$$

There are two more bits of algebra that we will need below. This is not particularly exciting, but later we will be happy to have them available.

$$\begin{aligned}s_b = {} & \frac{\mathrm{Cov}(p,q)\mathrm{Cov}(p,pq) - \mathrm{Cov}(q,pq)\mathrm{Var}(p)}{\big(\mathrm{Cov}(p,q)\big)^2 - \mathrm{Var}(p)\mathrm{Var}(q)} \\ & = \frac{\bar{p}(1-\bar{p})\big((1-r)\bar{p}+r\big)\big(r\bar{p}(1-\bar{p}) - \bar{p}(1-\bar{p})\big)}{\big(r\bar{p}(1-\bar{p})\big)^2 - \big(\bar{p}(1-\bar{p})\big)^2} = \frac{\big((1-r)\bar{p}+r\big)(r-1)}{(r^2-1)} \\ & = \frac{\big((1-r)\bar{p}+r\big)(1-r)}{1-r^2} = \frac{(1-r)\bar{p}+r}{1+r}\end{aligned}$$

Because $\mathrm{Cov}(q,pq) = \mathrm{Cov}(p,pq)$, we find that

$$s_c = \frac{\mathrm{Cov}(p,pq)\mathrm{Var}(q) - \mathrm{Cov}(p,q)\mathrm{Cov}(q,pq)}{\big(\mathrm{Cov}(p,q)\big)^2 - \mathrm{Var}(p)\mathrm{Var}(q)} = -s_b$$

B.2 Model 1

Model 1 is

$$w_i = \beta_0 + \beta_1 p_i + \varepsilon_i$$

In an infinite population, this means that the average fitness of "cooperators" and "defectors" is

$$\bar{w}_C = \beta_0 + \beta_1$$

$$\bar{w}_D = \beta_0$$

The overall average fitness is

$$\bar{w} = \bar{p}(\beta_0 + \beta_1) + (1 - \bar{p})\beta_0 = \beta_0 + \bar{p}\beta_1$$

This makes

$$\Delta\bar{p} = \frac{\bar{p}(\beta_0 + \beta_1)}{\bar{w}} - \bar{p} = \frac{\bar{p}(\beta_0 + \beta_1)}{\beta_0 + \bar{p}\beta_1} - \frac{\bar{p}(\beta_0 + \bar{p}\beta_1)}{\beta_0 + \bar{p}\beta_1} = \frac{\beta_1 \cdot \bar{p}(1 - \bar{p})}{\bar{w}}$$

This implies that

$$\bar{w}\Delta\bar{p} = \beta_1 \cdot \bar{p}(1 - \bar{p}) = \beta_1 \cdot \mathrm{Var}(p)$$

The Generalized Price equation implies that if we take statistical Model 1, then

$$\bar{w}\Delta\bar{p} = \hat{\beta}_1 \cdot \mathrm{Var}(p)$$

That implies that we do not have to do the actual minimizing of the sum of squared errors that we have done in Appendix A.1 to conclude that $\hat{\beta}_1 = \beta_1$ if we apply Price-like equation 1 to Model 1.

B.3 Model 2

Model 2 is

$$w_i = \gamma_{0,0} + \gamma_{1,0} p_i + \gamma_{0,1} q_i + \varepsilon_i$$

In an infinite population, this means that the average fitness of cooperators and defectors is

$$\bar{w}_C = \gamma_{0,0} + \gamma_{1,0} + \big((1 - r)\bar{p} + r\big)\gamma_{0,1}$$

$$\bar{w}_D = \gamma_{0,0} + (1 - r)\bar{p}\gamma_{0,1}$$

The overall average fitness is

$$\begin{aligned}\bar{w} = \bar{p}\big(\gamma_{0,0} + \gamma_{1,0} + \big((1-r)\bar{p} + r\big)\gamma_{0,1}\big) + (1-\bar{p})\big(\gamma_{0,0} + (1-r)\bar{p}\gamma_{0,1}\big) \\ = \gamma_{0,0} + \bar{p}\big(\gamma_{1,0} + \gamma_{0,1}\big)\end{aligned}$$

This makes

$$\begin{aligned}\Delta\bar{p} = \frac{\bar{p}\bar{w}_C}{\bar{w}} - \bar{p} = \frac{\bar{p}\big(\gamma_{0,0} + \gamma_{1,0} + \big((1-r)\bar{p} + r\big)\gamma_{0,1}\big)}{\gamma_{0,0} + \bar{p}\big(\gamma_{1,0} + \gamma_{0,1}\big)} - \frac{\bar{p}\left(\gamma_{0,0} + \bar{p}\big(\gamma_{1,0} + \gamma_{0,1}\big)\right)}{\gamma_{0,0} + \bar{p}\big(\gamma_{1,0} + \gamma_{0,1}\big)} \\ = \frac{\bar{p}(1-\bar{p})\big(\gamma_{1,0} + r\gamma_{0,1}\big)}{\bar{w}}\end{aligned}$$

This implies that

$$\bar{w}\Delta\bar{p} = \big(\gamma_{1,0} + r\gamma_{0,1}\big) \cdot \bar{p}(1-\bar{p}) = \big(\gamma_{1,0} + r\gamma_{0,1}\big) \cdot \mathrm{Var}(p)$$

The Generalized Price equation implies that if we take statistical Model 2, then

$$\bar{w}\Delta\bar{p} = \big(\hat{\gamma}_{1,0} + r\hat{\gamma}_{0,1}\big) \cdot \mathrm{Var}(p)$$

That implies that we do not have to do the actual minimizing of the sum of squared errors to conclude that $\hat{\gamma}_{1,0} = \gamma_{1,0}$ and $\hat{\gamma}_{0,1} = \gamma_{0,1}$ if we apply Price-like equation 2 to Model 2.

With $b = \gamma_{0,1}$ and $-c = \gamma_{1,0}$, here we of course recognize Hamilton's original rule.

The Generalized Price equation also implies that if we take statistical Model 1, then still

$$\bar{w}\Delta\bar{p} = \hat{\beta}_1 \cdot \mathrm{Var}(p)$$

That implies that also here, we do not have to do the actual minimizing of the sum of squared errors that we did in Appendix A.2 to conclude that, if we apply the Generalized Price equation using statistical Model 1 to Model 2, we get $\hat{\beta}_1 = \gamma_{1,0} + r\gamma_{0,1}$. For other mismatches, this shortcut does not work, and we will have to do the actual minimization (see also footnote 8).

B.4 Model 3

Model 3 is

$$w_i = \delta_{0,0} + \delta_{1,0}p_i + \delta_{0,1}q_i + \delta_{1,1}p_iq_i + \varepsilon_i$$

In an infinite population, this means that the average fitness of cooperators and defectors is

$$\bar{w}_C = \delta_{0,0} + \delta_{1,0} + \big((1-r)\bar{p} + r\big)\big(\delta_{0,1} + \delta_{1,1}\big)$$

$$\bar{w}_D = \delta_{0,0} + (1-r)\bar{p}\delta_{0,1}$$

The overall average fitness is

$$\bar{w} = \delta_{0,0} + \bar{p}\big(\delta_{0,1} + \delta_{1,0}\big) + \bar{p}\big((1-r)\bar{p} + r\big)\delta_{1,1}$$

This makes

$$\begin{aligned}
\Delta\bar{p} &= \frac{\bar{p}\bar{w}_C}{\bar{w}} - \bar{p} \\
&= \frac{\bar{p}\left(\delta_{0,0} + \delta_{1,0} + \big((1-r)\bar{p} + r\big)\big(\delta_{0,1} + \delta_{1,1}\big)\right)}{1 + \bar{p}\big(\delta_{0,1} + \delta_{1,0}\big) + \bar{p}\big((1-r)\bar{p} + r\big)\delta_{1,1}} \\
&\quad - \frac{\bar{p}\big(\delta_{0,0} + \bar{p}\big(\delta_{0,1} + \delta_{1,0}\big) + \bar{p}\big((1-r)\bar{p} + r\big)\delta_{1,1}\big)}{1 + \bar{p}\big(\delta_{0,1} + \delta_{1,0}\big) + \bar{p}\big((1-r)\bar{p} + r\big)\delta_{1,1}} \\
&= \frac{\delta_{1,0}\bar{p}(1-\bar{p}) + r\delta_{0,1}\bar{p}(1-\bar{p}) + \big((1-r)\bar{p} + r\big)\delta_{1,1}\bar{p}(1-\bar{p})}{\bar{w}}
\end{aligned}$$

This implies that

$$\begin{aligned}
\bar{w}\Delta\bar{p} &= \big(\delta_{1,0} + r\delta_{0,1} + \big((1-r)\bar{p} + r\big)\delta_{1,1}\big) \cdot \bar{p}(1-\bar{p}) \\
&= \big(\delta_{1,0} + r\delta_{0,1} + \big((1-r)\bar{p} + r\big)\delta_{1,1}\big) \cdot \mathrm{Var}(p)
\end{aligned}$$

The Generalized Price equation implies that if we take statistical Model 3, then

$$\bar{w}\Delta\bar{p} = \big(\hat{\delta}_{1,0} + r\hat{\delta}_{0,1} + \big((1-r)\bar{p} + r\big)\hat{\delta}_{1,1}\big) \cdot \mathrm{Var}(p)$$

That implies that we do not have to do the actual minimizing of the sum of squared errors to conclude that $\hat{\delta}_{1,0} = \delta_{1,0}$, $\hat{\delta}_{0,1} = \delta_{0,1}$, and $\hat{\delta}_{1,1} = \delta_{1,1}$ if we apply Price-like equation 3 to Model 3.

B.5 Price-like equation 2, applied to Model 3

The Generalized Price equation for Model 2 is

$$\bar{w}\Delta\bar{p} = \big(\hat{\gamma}_{1,0} + r\hat{\gamma}_{0,1}\big) \cdot \mathrm{Var}(p)$$

If we apply this to Model 3, then this does not allow for a straightforward computation of $\hat{\gamma}_{1,0}$ and $\hat{\gamma}_{0,1}$ without having to do the actual minimization of the sum of squared errors. The reason for that is that now have one equation, that is, we have $\hat{\gamma}_{1,0} + r\hat{\gamma}_{0,1} = \delta_{1,0} + r\delta_{0,1} + \big((1-r)\bar{p} + r\big)\delta_{1,1}$, and two variables, $\hat{\gamma}_{1,0}$ and $\hat{\gamma}_{0,1}$. What we can do, however, is just verify that the $\hat{\gamma}_{1,0}$ and $\hat{\gamma}_{0,1}$ that we found in Appendix A.3, work. There, we found that

$$\hat{\gamma}_{0,0} = \delta_{0,0} - (1-r)\bar{p}s_b\delta_{1,1}$$

$$\hat{\gamma}_{1,0} = \delta_{1,0} - s_c\delta_{1,1}$$

$$\hat{\gamma}_{0,1} = \delta_{0,1} + s_b\delta_{1,1}$$

with

$$s_b = \frac{(1-r)\bar{p} + r}{1+r}$$

$$s_c = -\frac{(1-r)\bar{p} + r}{1+r}$$

If we then fill those in, we get

$$\begin{aligned}\bar{w}\Delta\bar{p} = \left(\hat{\gamma}_{1,0} + r\hat{\gamma}_{0,1}\right)\cdot \mathrm{Var}(p) &= \left(\delta_{1,0} - s_c\delta_{1,1} + r\left(\delta_{0,1} + s_b\delta_{1,1}\right)\right)\cdot \mathrm{Var}(p)\\ &= \left(\delta_{1,0} + r\delta_{0,1} + (rs_b - s_c)\delta_{1,1}\right)\cdot \mathrm{Var}(p)\\ &= \left(\delta_{1,0} + r\delta_{0,1} + \left((1-r)\bar{p} + r\right)\delta_{1,1}\right)\cdot \mathrm{Var}(p)\end{aligned}$$

This is indeed what we saw above.

If the true model were Model 2, one would interpret $\hat{\gamma}_{1,0}$ as the effect of having the gene on oneself. The Generalized Price equation for Model 2 applied to Model 3 gives, as we saw above

$$\hat{\gamma}_{1,0} = \delta_{1,0} + \frac{(1-r)\bar{p} + r}{1+r}\delta_{1,1},$$

This is not a constant and cannot be interpreted as the effect of having the gene on oneself. We can search for a possible interpretation, by first calculating the *average* effect of having the gene, from the perspective of those with the gene, on their fitness. This would be $\delta_{1,0}$ plus $\delta_{1,1}$ times the probability that carriers are matched with another individual that has the gene;

$$\delta_{1,0} + \left((1-r)\bar{p} + r\right)\delta_{1,1}$$

We can also calculate the average effect of having the gene, from the perspective of those without the gene. This is $\delta_{1,0}$ plus $\delta_{1,1}$ times the probability that non-carriers are matched with an individual that has the gene;

$$\delta_{1,0} + (1-r)\bar{p}\delta_{1,1}$$

Finally, one could also average these average effects over carriers and non-carriers. That would give

$$\delta_{1,0} + (1-r)\bar{p}\delta_{1,1} + r\bar{p}\delta_{1,1} = \delta_{1,0} + \bar{p}\delta_{1,1}$$

We see that neither of these three matches the $\hat{\gamma}_{1,0}$ that we get from applying the Generalized Price equation for Model 2 to Model 3.

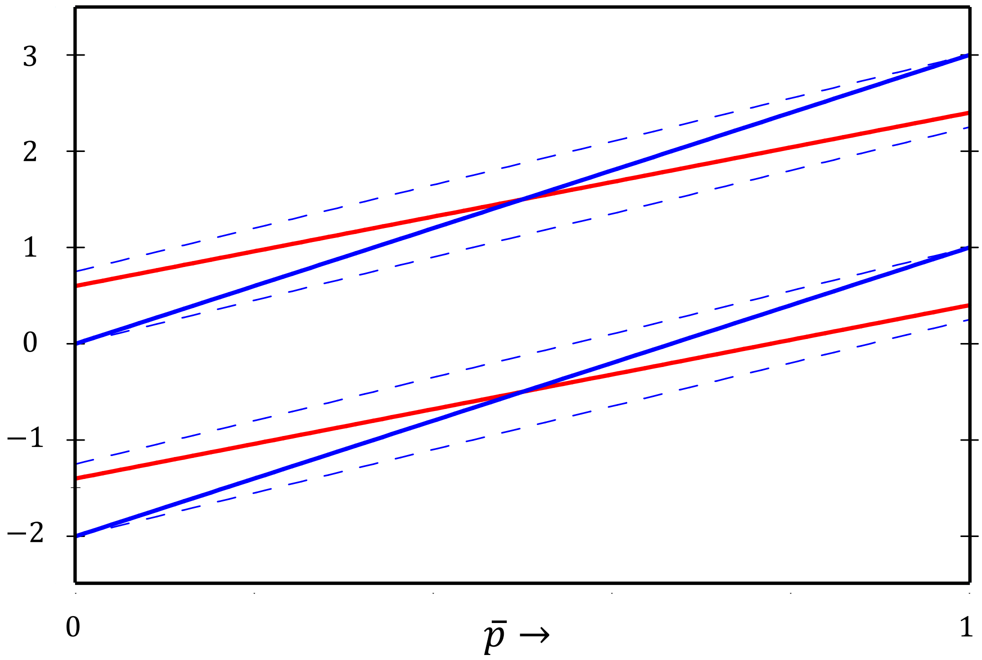

Fig. 6 | Applying the Generalized Price equation for Model 2 to Model 3. For this figure, we chose Model 3 with $\delta_{1,0} = -2$, $\delta_{1,1} = 0$, and $\delta_{1,1} = 3$. The red lines are $\hat{\gamma}_{1,0}$ (the lower one), and $\hat{\gamma}_{0,1}$ (the higher one), which, when applied to Model 2, would have been the effect on the individual itself and the effect on the partner. The average effect on the individual itself is the lower unbroken blue line, which is accompanied by the average effect for the defectors (the dashed line below it) and the average effect for cooperators (the dashed line immediately above it). The average effect on the partner is the higher unbroken blue line, which is accompanied by the average effect for the defectors (the dashed line immediately below it) and the average effect for cooperators (the dashed line above it).

If the true model were Model 2, one would interpret $\hat{\gamma}_{0,1}$ as the effect of having the gene on one's partner. The Generalized Price equation for Model 2 applied to Model 3 gives

$$\hat{\gamma}_{0,1} = \delta_{0,1} + \frac{(1-r)\bar{p} + r}{1+r}\delta_{1,1},$$

This is not a constant either and cannot be interpreted as the effect of having the gene on the interaction partner. We can search for a possible interpretation, by first calculating the *average* effect of having the gene, from the perspective of those with the gene, on their partner's fitness. This is $\delta_{0,1}$ plus $\delta_{1,1}$ times the probability that carriers are matched with another individual that has the gene.

$$\delta_{0,1} + \big((1-r)\bar{p} + r\big)\delta_{1,1}$$

We can also compare it with the average effect of having the gene, from the perspective of those without the gene, on their partner's fitness. This is $\delta_{0,1}$ plus $\delta_{1,1}$ times the probability that non-carriers are matched with an individual that has the gene.

$$\delta_{0,1} + (1 - r)\bar{p}\delta_{1,1}$$

Finally, one could also average these average effects over carriers and non-carriers. That would give

$$\delta_{0,1} + (1 - r)\bar{p}\delta_{1,1} + r\bar{p}\delta_{1,1} = \delta_{0,1} + \bar{p}\delta_{1,1}$$

We see that also here, neither of these three matches the $\hat{\gamma}_{0,1}$ that we get from applying the Generalized Price equation for Model 2 to Model 3.

Fig. 6 above illustrates this and visualizes that trying to understand the working of Model 3 through applying the Generalized Price equation for Model 2 on it, is misleading. The values for $\hat{\gamma}_{1,0}$ and $\hat{\gamma}_{0,1}$ that this results in do not coincide with the average effect on the individual itself and the average effect on the partner.

References

[1] G. R. Price, 'Selection and Covariance', *Nature*, vol. 227, no. 5257, pp. 520–521, Aug. 1970, doi: 10.1038/227520a0.

[2] M. van Veelen, 'On the use of the Price equation', *J Theor Biol*, vol. 237, no. 4, 2005, doi: 10.1016/j.jtbi.2005.04.026.

[3] M. van Veelen, 'The problem with the Price equation', *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 375, no. 1797, p. 20190355, Apr. 2020, doi: 10.1098/rstb.2019.0355.

[4] M. van Veelen, J. García, M. W. Sabelis, and M. Egas, 'Call for a return to rigour in models', *Nature*, vol. 467, no. 7316, 2010, doi: 10.1038/467661d.

[5] M. van Veelen, J. García, M. W. Sabelis, and M. Egas, 'Group selection and inclusive fitness are not equivalent; the Price equation vs. models and statistics', *J Theor Biol*, vol. 299, 2012, doi: 10.1016/j.jtbi.2011.07.025.

[6] G. R. Price, 'Extension of covariance selection mathematics', *Ann Hum Genet*, vol. 35, no. 4, pp. 485-490., 1972.

[7] A. Grafen, 'A geometric view of relatedness', *Oxford surveys in evolutionary biology*, vol. 2, no. 2, pp. 28–89, 1985.

[8] A. Grafen, 'Developments of the Price equation and natural selection under uncertainty', *Proc R Soc Lond B Biol Sci*, vol. 267, no. 1449, pp. 1223–1227, Jun. 2000, doi: 10.1098/rspb.2000.1131.

[9] P. D. Taylor, 'Evolutionary stability in one-parameter models under weak selection', *Theor Popul Biol*, vol. 36, no. 2, pp. 125–143, Oct. 1989, doi: 10.1016/0040-5809(89)90025-7.

[10] P. D. Taylor and S. A. Frank, 'How to Make a Kin Selection Model', *J Theor Biol*, vol. 180, no. 1, pp. 27–37, May 1996, doi: 10.1006/jtbi.1996.0075.

[11] S. A. Frank, *Foundations of social evolution*. Princeton: Princeton University Press, 1998.

[12] S. Rice, *Evolutionary theory: mathematical and conceptual foundations*. Sinauer Associates, 2004.

[13] S. Okasha, *Evolution and the levels of selection*. Clarendon Press, 2006.

[14] A. Grafen, ‘Optimization of inclusive fitness’, *J Theor Biol*, vol. 238, no. 3, pp. 541–563, Feb. 2006, doi: 10.1016/j.jtbi.2005.06.009.

[15] A. Gardner, ‘The Price equation’, *Current Biology*, vol. 18, no. 5, pp. R198–R202, Mar. 2008, doi: 10.1016/j.cub.2008.01.005.

[16] A. Gardner, S. A. West, and G. Wild, ‘The genetical theory of kin selection’, *J Evol Biol*, vol. 24, no. 5, pp. 1020–1043, May 2011, doi: 10.1111/j.1420-9101.2011.02236.x.

[17] S. A. Frank, ‘Natural selection. IV. The Price equation*’, *J Evol Biol*, vol. 25, no. 6, pp. 1002–1019, Jun. 2012, doi: 10.1111/j.1420-9101.2012.02498.x.

[18] F. Rousset, ‘Regression, least squares, and the general version of inclusive fitness’, *Evolution (N Y)*, vol. 69, no. 11, pp. 2963–2970, Nov. 2015, doi: 10.1111/evo.12791.

[19] J. A. R. Marshall, *Social evolution and inclusive fitness theory: an introduction*. Princeton University Press, 2015.

[20] D. C. Queller, ‘Fundamental Theorems of Evolution’, *Am Nat*, vol. 189, no. 4, pp. 345–353, Apr. 2017, doi: 10.1086/690937.

[21] V. J. Luque, ‘One equation to rule them all: a philosophical analysis of the Price equation’, *Biol Philos*, vol. 32, no. 1, pp. 97–125, Jan. 2017, doi: 10.1007/s10539-016-9538-y.

[22] M. van Veelen, ‘Can Hamilton’s rule be violated?’, *Elife*, vol. 7, Oct. 2018, doi: 10.7554/eLife.41901.

[23] M. van Veelen, ‘The group selection–inclusive fitness equivalence claim: not true and not relevant’, *Evol Hum Sci*, vol. 2, p. e11, Apr. 2020, doi: 10.1017/ehs.2020.9.

[24] M. van Veelen, B. Allen, M. Hoffman, B. Simon, and C. Veller, ‘Hamilton’s rule’, *J Theor Biol*, vol. 414, 2017, doi: 10.1016/j.jtbi.2016.08.019.

[25] B. Allen, M. A. Nowak, and E. O. Wilson, ‘Limitations of inclusive fitness’, *Proceedings of the National Academy of Sciences*, vol. 110, no. 50, pp. 20135–20139, Dec. 2013, doi: 10.1073/pnas.1317588110.

[26] M. A. Nowak, A. McAvoy, B. Allen, and E. O. Wilson, ‘The general form of Hamilton’s rule makes no predictions and cannot be tested empirically’, *Proceedings of the National Academy of Sciences*, vol. 114, no. 22, pp. 5665–5670, May 2017, doi: 10.1073/pnas.1701805114.

[27] C. Darwin, *On the origin of species*. 1859.

[28] R. Dawkins, *The selfish gene*. 1976.

[29] M. A. Nowak, C. E. Tarnita, and E. O. Wilson, ‘The evolution of eusociality’, *Nature*, vol. 466, no. 7310, pp. 1057–1062, Aug. 2010, doi: 10.1038/nature09205.

[30] P. Abbot *et al.*, ‘Inclusive fitness theory and eusociality’, *Nature*, vol. 471, no. 7339, pp. E1–E4, Mar. 2011, doi: 10.1038/nature09831.

[31] W. D. Hamilton, ‘The genetical evolution of social behaviour. I’, *J Theor Biol*, vol. 7, no. 1, pp. 1–16, Jul. 1964, doi: 10.1016/0022-5193(64)90038-4.

[32] M. van Veelen, ‘Hamilton’s missing link’, *J Theor Biol*, vol. 246, no. 3, 2007, doi: 10.1016/j.jtbi.2007.01.001.

[33] W. D. Hamilton, ‘Selfish and Spiteful Behaviour in an Evolutionary Model’, *Nature*, vol. 228, no. 5277, pp. 1218–1220, Dec. 1970, doi: 10.1038/2281218a0.

[34] W. D. Hamilton, ‘Innate social aptitudes of man: an approach from evolutionary genetics’, in *Narrow roads of gene land. Vol. 1: Evolution of social behaviour*, R. Fox, Ed., 1975, pp. 315–352.

[35] D. C. Queller, ‘A general model for kin selection’, *Evolution (N Y)*, vol. 46, no. 2, pp. 376–380, Apr. 1992, doi: 10.1111/j.1558-5646.1992.tb02045.x.

[36] T. Kay, L. Keller, and L. Lehmann, ‘The evolution of altruism and the serial rediscovery of the role of relatedness’, *Proceedings of the National Academy of*

Sciences, vol. 117, no. 46, pp. 28894–28898, Nov. 2020, doi: 10.1073/pnas.2013596117.

[37] P. D. Taylor, T. Day, and G. Wild, ‘Evolution of cooperation in a finite homogeneous graph’, *Nature*, vol. 447, no. 7143, pp. 469–472, May 2007, doi: 10.1038/nature05784.

[38] A. Akdeniz and M. van Veelen, ‘The cancellation effect at the group level’, *Evolution (N Y)*, vol. 74, no. 7, pp. 1246–1254, Jul. 2020, doi: 10.1111/evo.13995.

[39] D. C. Queller, ‘Kinship, reciprocity and synergism in the evolution of social behaviour’, *Nature*, vol. 318, no. 6044, pp. 366–367, Nov. 1985, doi: 10.1038/318366a0.

[40] A. Grafen, ‘News and Views. Evolutionary theory: Hamilton’s rule OK’, *Nature*, vol. 318, no. 6044, pp. 310–311, Nov. 1985, doi: 10.1038/318310a0.

[41] C. Matessi and S. Karlin, ‘On the evolution of altruism by kin selection’, *Proceedings of the National Academy of Sciences*, vol. 81, no. 6, pp. 1754–1758, Mar. 1984, doi: 10.1073/pnas.81.6.1754.

[42] M. van Veelen, ‘Group selection, kin selection, altruism and cooperation: When inclusive fitness is right and when it can be wrong’, *J Theor Biol*, vol. 259, no. 3, 2009, doi: 10.1016/j.jtbi.2009.04.019.