J. theor. Biol. (1990) **144**, 517–546

Biological Signals as Handicaps

ALAN GRAFEN†

Animal Behaviour Research Group, Zoology Department, South Parks Road, Oxford OX1 3PS, U.K.

(*Received on* 25 *August* 1989, *Accepted on* 13 *October* 1989)

An ESS model of Zahavi's handicap principle is constructed. This allows a formal exposition of how the handicap principle works, and shows that its essential elements are strategic. The handicap model is about signalling, and it is proved under fairly general conditions that if the handicap principle's conditions are met, then an evolutionarily stable signalling equilibrium exists in a biological signalling system, and that any signalling equilibrium satisfies the conditions of the handicap principle. Zahavi's major claims for the handicap principle are thus vindicated. The place of cheating is discussed in view of the honesty that follows from the handicap principle. Parallel signalling models in economics are discussed. Interpretations of the handicap principle are compared. The models are not fully explicit about how females use information about male quality, and, less seriously, have no genetics. A companion paper remedies both defects in a model of the handicap principle at work in sexual selection.

1. Introduction

The application of adaptationist principles to animal communication has produced two apparently conflicting traditions. Dawkins & Krebs (1978; Krebs & Dawkins, 1984), stress that animals cheat and manipulate when they communicate. On the other hand, Zahavi (1975, 1977, 1987) concludes that biological signals must be honest, and goes on to draw many conclusions from this. Here I resolve this conflict, partly by providing mathematical models. These game theory models will show Zahavi's handicap principle at work: each organism maximizes its fitness, and signals are honest. The models clarify Zahavi's handicap principle, and show him to have been substantially correct in his claims for its importance and scope.

The only previous model of biological signalling of which I am aware is that of Enquist (1985) who showed that, contrary to a (still) popular belief, players in evolutionary games could communicate information about their intentions in an evolutionarily stable way. Enquist established criteria for what counts as a signal, and these criteria and the main conclusions of his models are discussed below.

This and a companion paper (this phrase will refer to Grafen, 1990) present a number of models, all based on the handicap principle of Zahavi (1975, 1977). The present paper uses ESS models to consider signals, and affirms Zahavi's (1987) claim that natural selection on a wide class of signals necessarily incurs waste in

†† Present address: Department of Plant Sciences, South Parks Road, Oxford OX1 3RA, U.K.

accordance with the handicap principle. The companion paper uses a population genetics model to defend the centrality of the handicap principle in sexual selection.

To avoid unnecessarily abstract discussion, the present paper begins with one ESS model of signalling in section 2, and another in section 3. Because concrete females and males are easier to understand than abstract signallers and receivers, the models of the first two sections are phrased in terms of sexual selection and mate choice. Section 4 considers other applications of these same models. Section 5 considers limitations of these models as general signal models, and goes on to discuss what can nevertheless be concluded from them. Section 6 explains what happened to cheating in the outburst of honesty created by Zahavi's handicap principle, while section 7 poses the question: what are signals? Parallel models in economics are discussed in section 8. Section 9 is a brief review of previous models of the handicap principle, putting them in the context of the understanding of the handicap principle developed in earlier sections. Concluding remarks are made in section 10.

2. An ESS Model of Strategic Choice Handicaps

The essential elements that define the signalling systems to which the model applies are

(i) males vary in some quantity of interest to females, which females cannot observe, but which, if they could, they would be selected to use in mate choice. This variable will be denoted q for true quality, and it will be assumed that the higher a mate's q is the better for the female. A male cannot alter his own value of q.

(ii) males vary in some observable quantity, which will be denoted a for advertizing. The level of advertizing of a male may depend on his level of q. Males can alter their value of a, and a male strategy is a function $A(q)$ which determines a level of advertizing corresponding to each true quality.

(iii) females use the observed value of a to infer a male's value of q. The inferred value of q for a male will be denoted p for perceived value. A female's strategy is a function $P(a)$ which determines the perceived value of a male with each possible level of advertizing. It is assumed that males with a higher perceived value are fitter, as a consequence of females' response to them, than otherwise similar males.

q, a and p will all be taken as real numbers for the remainder of the paper, although some of the formalism makes sense more generally and it then suggests further topics in strategic handicap theory.

The fitness of a male depends on his true quality, his advertizing level and his perceived value, and will be denoted $w(a, p, q)$. The fitness of a female will be assumed to depend on the discrepancies in her perception of males' true qualities. Suppose $D(q, p)$ is the loss in fitness to a female assessing as p a male with true quality q. We assume that

$$\begin{aligned} D(q,p) &> 0 \quad q \neq p \\ &= 0 \quad q = p, \end{aligned}$$

so that the female does best to assess correctly. Then we assume a female's fitness is reduced by the average D over all males in the population, and write $G(q)$ for the cumulative frequency distribution of q among males. That average can be written as

$$\int D[q, P(a)]\, \mathrm{d}G(q),$$

where a is taken to vary with q.

When considering if a pair of strategies $A^*(q)$, $P^*(a)$ is evolutionarily stable, we can assume that they are universal in the population. Hence we can write the conditions that the pair is evolutionarily stable as

$$w[A^*(q), P^*(A^*(q)), q] \geq w[a, P^*(a), q] \quad \text{for all } a, q. \tag{1}$$

$$\int D[q, P^*(A^*(q))]\, \mathrm{d}G(q) \leq \int D[q, P(A^*(q))]\, \mathrm{d}G(q) \text{ for all functions } P(a).$$

The first inequality states that $A^*(q)$ is the best level of advertizing for a male of quality q. The second inequality states that P^* is the best possible female assessment rule. In this form, the evolutionary game is well specified with three arbitrary functions w, D and G. I shall make the assumption that the set of points of increase of G is an interval, which means that there are no gaps in the distribution of quality. We shall see that the precise forms of D and G do not matter, and that we can reach all the important conclusions by making only a few assumptions about partial derivatives of w.

This is an ESS model (Evolutionarily Stable Strategy—see Maynard Smith, 1982) of a signalling system. We will now use it. First to show that under weak conditions on the function $w(a, p, q)$, an ESS exists which exhibits the features Zahavi associated with the handicap principle. Then it will be used to prove the main handicap result, which is a general claim about what must be true of a whole class of signalling systems. In addition, the ESS model is extremely useful as a "stencil" for constructing the biological parts of the population genetic model in the companion paper.

2.1. THE EXISTENCE OF HANDICAP EQUILIBRIA

Under mild conditions on $w(a, p, q)$, an ESS exists. These conditions are fairly natural mathematical representations of Zahavi's conclusions about signalling systems in general. In order to be reasonably rigorous, some of the assumptions are rather technical.

We need to assume that $w(a, p, q)$ is continuous. The notation w_1, w_2, w_3 will be used to denote the partial derivatives with respect to the first, second and third arguments, and multiple subscripts will represent corresponding higher order derivatives. We assume that w_1, w_2, and w_3 exist. It is an essential part of the handicap principle that advertizing is costly, so that w_1 is negative. Equally, if males strive to improve females' assessment of their trait, then their fitness must increase when females' perception of their trait increases. Hence w_2 is positive. It must be that

better males do better by advertizing more, and the condition that ensures this is that

$$\frac{w_1(a, p, q)}{w_2(a, p, q)}$$

is strictly increasing in q. This means that the ratio of the marginal cost of advertizing to the marginal advantage of improved assessment by females must be an increasing function of quality. If we assume that w_{13} and w_{23} exist, then this condition will be satisfied when $w_{23} \geq 0$ and $w_{13} > 0$. If $w_{23} = 0$, then this reduces to the condition that the marginal cost of advertizing (w_1) should be greater for worse males, i.e. $w_{13} \geq 0$. $w_{23} \geq 0$ means that the advantage gained by a male through an improvement in females' assessment of him is at least as great for better males as for worse males. These assumptions follow directly from one of Zahavi's descriptions of the handicap principle (Zahavi, 1977: section 2). For technical reasons we assume that the ratio w_1/w_2 is defined. The set of values of q for which $G(q)$ is increasing is assumed to be an interval on the real line, $[q_{\min}, q_{\max}]$, where $q_{\min}$ is finite. Also, I suppose that there is a finite minimum level of advertizing, $a_{\min}$.

Calculations in Appendix 2 now show that we can define functions P^* and A^* as follows

$$\begin{aligned} P^*(a_{\min}) &= q_{\min} \\ P^{*\prime}(a) &= -\frac{w_1[a, P^*(a), P^*(a)]}{w_2[a, P^*(a), P^*(a)]} \\ P^*[A^*(q)] &= q \end{aligned} \tag{2}$$

and that A^*, P^* is an ESS pair of strategies. In other words, if all males are playing A^* and all females are playing P^*, then no male can do better than by playing A^*, and no female can do better than by playing P^*.

With this constructive result on the existence of an ESS, we can draw general conclusions about the form of this ESS, and also choose example functions for $w(a, p, q)$ and calculate what the ESS is. $A^*(q)$ is an increasing function of q, so that males with higher q advertize more. Despite having a free strategic choice of a, males choose an advertizing level which can be used to pinpoint their true quality. $P^*[A^*(q)] = q$, so that females correctly infer a male's quality from his advertizing level. Finally, the net effect of advertizing and female choice is that males with higher q end up with higher w. Higher quality males therefore advertize more, but these costs are more than compensated for by the consequences of advertizing on female preference.

In this situation the signal is acting as a handicap. All males voluntarily pay higher advertizing costs than they need, better males advertize more, and females use advertizing as a reliable guide to quality. If advertizing were not costly, then the signal could not operate in this way; nor if it were equally costly to good and bad males. The cost of the signal is therefore essential to its operation. It therefore makes sense to say that the reason males signal in this way is because it is costly. The signal is selected because it reduces the fitness of its bearer. More precisely, it

reduces one component of the bearer's fitness, and the over-compensating increase in the other component depends on the interpretation by females of the signal.

It might be thought that a sensible way to show that handicaps are costly is to show a negative correlation between size of the handicap and some component of fitness. This is unlikely to be right, because it neglects the fact of inter-individual variation in quality, one of the prime requirements of the signal theory. What correlations across males does one expect to observe in a population at a signalling equilibrium? The results about the ESS just given imply that the level of the handicap, quality, attractiveness to females and net fitness would all be strictly and positively correlated with each other.

2.2. THE MAIN HANDICAP RESULT

In this section, the argument of the previous section is reversed to derive a result which is much closer to the spirit of Zahavi's main argument, which is that one can conclude from the evolutionary stability of signals that they are honest, costly and costly in a way that relates to the true quality revealed. Translated into formal terms, this statement becomes:

If A^*, P^* is an ESS, $w_2 > 0$ and $A^*(q)$ is increasing, then

(a) $P^*[A^*(q)] = q \qquad \forall q$ (honesty)

(b) $w_1 < 0$ (cost)

(c) $\dfrac{w_1(a, p, q)}{w_2(a, p, q)}$ is strictly increasing in q near the path $[A^*(q), q, q]$ (costlier for worse males).

This result is proved in Appendix 3. Its importance is that handicaps are not just one quirky possibility. If we see a character which does signal quality, then it must be a handicap. The handicap principle lies at the heart of evolutionary signalling, and must therefore play a major rôle in our understanding of it.

A rough verbal version of the mathematical argument runs as follows. The first thing is to show that evolutionarily stable signals must be honest. Call a signal reliable if it can be used to determine the true quality of a male. At equilibrium, signals must be reliable, otherwise females would not use them. If signals are reliable, then at ESS the receivers will have adjusted their assessment rule so that they determine correctly the true quality of a male from his level of advertizing. Therefore at ESS, signals are honest. Now it is easy to show the other two parts. Males have a free strategic choice of their level of advertizing. A male who advertizes more will experience an increase in his fitness through the increased assessment females will make of his true quality q. If a male's level of advertizing is evolutionarily stable, then it cannot pay him to advertize a little bit more. Therefore advertizing more must be costly, and this cost must more than balance the gain which would be achieved by advertizing more. But advertizing is also known to be honest, so that a better male advertizes more than a worse one. Each male's level of advertizing is

evolutionarily stable, yet the better male advertizes more. It follows that the marginal cost of advertizing must be lower for the better male. This is the last of Zahavi's inferences.

2.3. AN EXAMPLE

A particular function for $w(a, p, q)$ is now assumed to show how strategic choice signalling works in an example. Suppose that

$$w(a, p, q) = p^r q^a.$$

This function can be interpreted as males beginning with a quality q which is just their viability. They raise this to a power a by advertizing (the minimum level of a is assumed to be 1) to produce their net viability. Given that a male does survive, his fitness is proportional to his number of mates, which is assumed to be p^r, where r is some positive constant. If $r = 1$, then a male has a number of mates proportional to his perceived original viability. By increasing or decreasing r we can tune the strength of the effect of female choice on male fitness. Roughly it is the same as if females choose the best male out of a random selection of $r+1$ males. If r is large, then only the very best males gain many mates—if r is low, then males are much more similar in their number of mates.

The ESS functions A and P can be worked out to be

$$P(a) = q_0^{\exp[-(a-a_0)/r]}$$

$$A(q) = a_0 - r.\ln\left(\frac{\ln(q)}{\ln(q_0)}\right).$$

We can also work out the net fitness of a male with true quality q. It turns out to be

$$w[A(q), P(A(q)), q] = q^r q^{a_0} q^{-r.\ln(\ln(q)/\ln(q_0))}.$$

Figure 1 shows $A(q)$ as a function of q for two different values of r. Figure 2 shows net viability as a function of q. Figure 3 shows net fitness as a function of q. Better

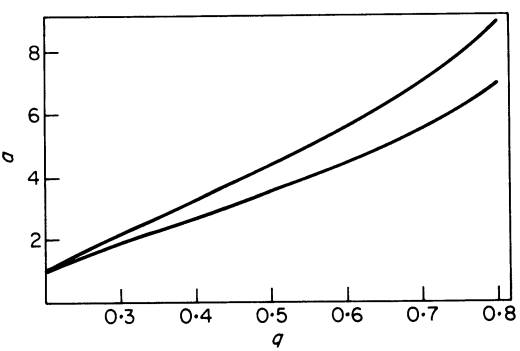

FIG. 1. This shows the optimal advertizing rule (a) as a function of quality (q) in the example of section 2.3 with $q_0 = 0{\cdot}2$, and $r = 4$ for upper function, $r = 3$ for the lower.

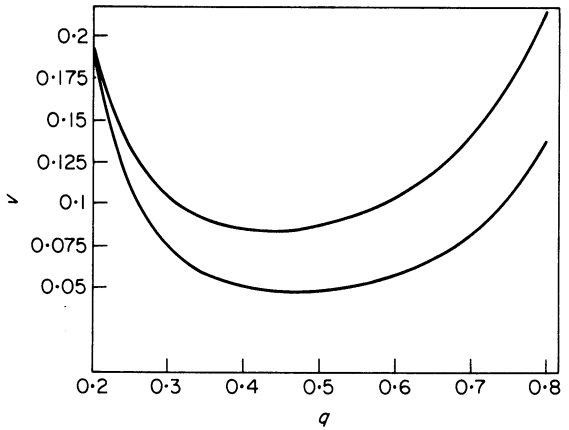

FIG. 2. The net viability of males (v) as a function of their quality (q) in the same example as Fig. 1. $r = 3$ for the upper curve, and $r = 4$ for the lower, showing that fussier females reduce the net viability of males. The worst males at $q = 0{\cdot}2$ advertize with $a = 1$, so their net viability is $q^1 = 0{\cdot}2$. The net viability is lowest for intermediate males. Net fitness, on the other hand, must increase with quality on very general grounds, as shown for this example in Fig. 3.

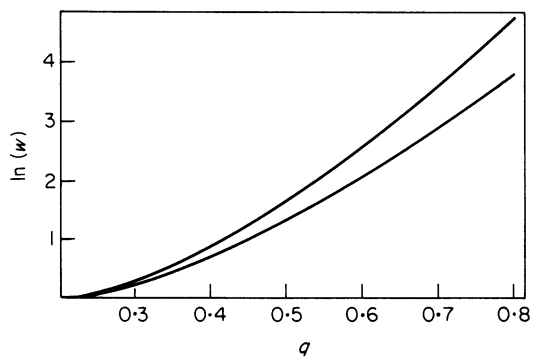

FIG. 3. This shows the natural logarithm of net fitness relative to the fitness of the worst male [ln (w)] as a function of quality (q) in the example of section 2.3, in the same example as Figs 1 and 2. The ratio of the fitnesses of the best to the worst male is 43·8865 for $r = 3$ and 112·957 for $r = 4$.

males advertize more, but there is no consistent effect on net viability, which decreases and then increases with q. Net fitness, on the other hand, must increase with q.

Two points emerge from this example. The first is the effect of the parameter r. r is the power to which p is raised in fitness, and corresponds roughly to what would result from a best of $r+1$ selection by females. The higher the value of r, the greater the level of advertizing in equilibrium, the more strongly the net fitness depends on q, and the higher the average value of q obtained by the females. r measures the extent of scrutiny exerted by females, and naturally this influences the rigour of the test which males choose to undergo. Figures 1 to 3 all provide a

comparison of the cases $r=3$ and $r=4$. Figure 4 shows the logarithm of the ratios of the highest and lowest fitnesses as a functon of r, which turns out to be a straight line.

The second point is that something else, rather more unexpected, can have a great effect on the level of advertizing. It is q_0, the viability of the worst male. In the example illustrated in the figures, q_0 has not been set equal to zero. This is because as q_0 approaches zero, the general level of advertizing increases indefinitely. Why is this? A good male advertizes to show that he is better than a slightly worse male, and has to advertize more than him to such an extent that it is not worthwhile for the slightly worse male to pretend to be the slightly better male. As there are more and more levels below an individual, they use up more and more advertizing space. Consequently a male of given quality has to advertize more and more, and that is one reason why as q_0 gets smaller, the general level of advertizing increases indefinitely. Another effect in this particular model is that at low values of q, the trade-off between advertizing and viability means that bad males will advertize a lot to distinguish themselves from slightly worse males. Low-life competition uses up a lot of advertizing space, and the respectably viable males at the top have no choice but to advertize even more. So the force of simply increasing the range of viabilities, and the force of the recklessness of advertizing expenditure of low viability males, mean that if the least viable male has a very low viability, then the top males will have very high levels of advertizing. Their net viability can be reduced as much as we like by making q_0 low enough. It might be worth considering in comparative studies of the exaggeration of sexually selected characters, how bad the worst males are.

This example has illustrated some features, but the main point of the models is the general solution which allows us to claim that all signalling that satisfies certain broad assumptions must also satisfy the Zahavian principles of honesty, cost and differential costs by quality.

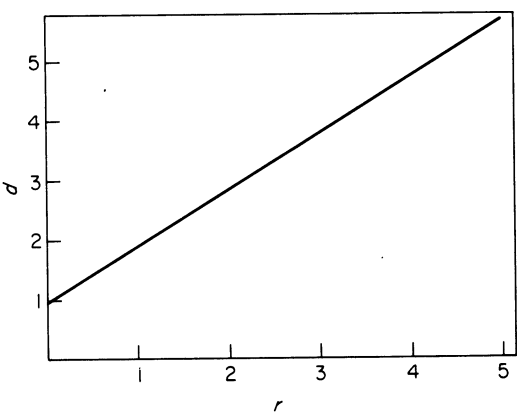

FIG. 4. The logarithm of the ratio of maximal to minimal net fitnesses (d) as a function of r, with $q_0=0{\cdot}2$ and $q_1=0{\cdot}8$. The function is linear in r.

3. A Modified Model in Which Females Pay the Cost of the Handicap

In the ESS model presented earlier, the females pay no cost for mating with a handicapped male. The disadvantage incurred by sons was one of the objections of Maynard Smith (1976) to Zahavi's (1975) original paper. Although our primary interest is not in sexual selection, it is of interest to consider what would happen if in disadvantaging himself by advertizing, a male also reduced his value to females. Could handicaps still evolve? We now construct a model in which instead of being interested in the male's original viability, we assume females are selected to mate with a male with the highest net viability, after advertizing has taken its toll. If the handicap principle still operates, then the females will be paying the cost of the handicap in full and still be using the handicap as a stable signal of quality.

To construct this model, we define net viability as $v(a, q)$, depending on true quality and advertizing. Then let net male fitness be $w(p, v)$. Because females pay the cost of the handicap, we assume a female's fitness depends on her mate's net viability, and not her mate's original viability. Thus p will be taken as a female's estimate of a male's net viability. Females in this model are selected to minimize the discrepancy between p and v, not between p and q. The conditions for A^*, P^* to be an ESS pair of strategies are therefore

$$w[v(A^*(q), P^*(A^*(q))), q] \geq w[v(a, P^*(a)), q] \quad \text{for all } q, a$$

$$\int D[v(A^*(q), q), P^*(A^*(q))]\, \mathrm{d}G(q)$$

$$\leq \int D[v(A^*(q), q), P(A^*(q))]\, \mathrm{d}G(q) \quad \text{for all functions } P. \tag{3}$$

Calculations in Appendix 4 show that the functions P^* and A^* can be defined using an intermediate function Q thus

$$Q(a_{\min}) = q_{\min}$$

$$Q'(a) = -\frac{v_1[a, Q(a)]}{v_2[a, Q(a)]}\left(1 + \frac{w_2[v(a, Q(a)), v(a, Q(a))]}{w_1[v(a, Q(a)), v(a, Q(a))]}\right) \tag{4}$$

$$A^*[Q(a)] = a$$

$$P^*(a) = v[a, Q(a)].$$

The pair A^*, P^* solves the ESS conditions and is an ESS. This ESS has less extreme advertizing than the case where females did not pay the cost of the handicap, as we would expect. The main features are unaltered. Male net fitness and advertizing level are still monotonically related to quality.

We can examine an example analogous to that used earlier, in which we choose the functons $w = p^r v$ and $v = q^a$. Then the solutions are:

$$A(q) = a_0\left(\frac{\ln q}{\ln q_0}\right)^{-r/(r+1)}$$

$$P(a) = \ln(q_0) a_0^{1+1/r} a^{-1/r},$$

and the net fitness function is given by

$$\ln(w) = -a_0(r+1)\ln(q_0)\left(\frac{\ln q}{\ln q_0}\right)^{1/(r+1)}.$$

This solution has all the main properties of the previous model. Unlimited exaggeration as q_0 goes to zero, and as r goes to infinity. One difference is that here, net viability of males does increase with quality. The reason is that females are interested in net viability, and they must get what they want in a stable signalling system.

The case where females pay the full cost of the handicap therefore leads to a lower level of advertizing, but leaves unchanged all the essential features of the ESS model of strategic choice handicaps. This model illustrates the inferiority of equations to words, because the equations must be reworked to get this answer. The verbal argument leading to the fundamental handicap principle applies in just the same way to this case as to the first.

4. Other Interpretations of the Models

For reasons of convenience, the previous two sections have used mate choice sexual selection as examples. The purpose of this section is to show that the same formalism covers other kinds of signalling. This range is important in assessing how generally we can apply conclusions based on those models.

Consider first the other traditional type of sexual selection, male-male competition. The model of section 2 can be applied in the following way. A male red deer holding a harem faces challenges from a succession of males that do not hold harems. Such a challenger must choose how strongly to roar in advertizing his strength. His ability to fight, or energy reserves for lasting out an exhausting engagement, can be taken as his quality, q of section 2. His roaring level can be taken as a, advertizing, and the harem master's assessment of the challenger's energy reserves is p. It is reasonable that the challenger is better off if he has higher energy reserves ($w_3 > 0$), and the higher the harem master's assessment of his reserves ($w_2 > 0$). Also that roaring itself reduces his fitness ($w_1 < 0$), and that roaring is more expensive in fitness for males with lower reserves ($w_{13} > 0$). Finally, it is reasonable that the gain in fitness from a better assessment by the harem master of a challenger's reserves is at least as great for a challenger with higher than lower reserves ($w_{23} > 0$).

The harem master's side of the model works out just as simply. In assessing a challenger's energy reserves from his roaring level, it pays a harem master to get it right, and he loses by either under- or over-estimation. Hence $D(p, q) = 0$ if $p = q$, and $D(p, q) > 0$ otherwise. This model therefore predicts there should be a graded response in which challengers with more energy reserves should roar at a higher level, and harem masters interpret this signal and treat the challenger appropriately.

This example extends to any kind of fighting. Our next example is begging nestlings. Here a nestling is competing with its nestmates for food supplied by its

parents. We apply the model by supposing that the parent wishes to feed the best growing chicks, so as not to waste food on the sick. The a is the energeticness of begging, p is the parent's assessment of the chick's state of growth, and q is the chick's true state of growth. $w_1 < 0$, as begging wastes energy. $w_2 > 0$, as being fed more is good for the chick. $w_{13} > 0$, as it is plausible that spending a given amount of energy begging harms the fitness of smaller chicks more. $w_{23} \geq 0$ would mean that the fitness gained by a marginal improvement in the parent's assessment of a chick is at least as great for big as for small chicks. This assumption would cease to be true if a chick were fed more than it could digest, and this fact may limit the wastefulness of begging at times when parents can supply an abundance of food.

As a final Zahavian example, consider one member of an antelope herd that is fleeing a lion. Consider the height of its jumps as a signal to the lion about the athletic ability of the antelope. The height of jumps is then a. The true athletic ability is q, and the athletic ability of the antelope as perceived by the lion is p. $w_1 < 0$, as jumping too high reduces the speed at which the antelope moves away from the lion. $w_2 > 0$, as the more athletic the lion perceives the antelope to be, the more likely she is to choose to pursue another member of the antelope herd, or alternatively to give up the chase. $w_{13} > 0$ means that it is more dangerous for a less athletic antelope to jump high, which is reasonable because there is a chance that the lion will attack. $w_{23} \geq 0$ means that the gain in fitness through being perceived as more athletic is at least as great for more as for less athletic antelope.

The first point of this section is that the models of sections 2 and 3 are models that can be applied to many kinds of signalling, and are not restricted to the female choice type of sexual selection. It is *not* a point of this section that these are the correct explanations of roaring in red deer, nestling begging or stotting. Zoology is not an armchair subject in which these matters of fact can be decided in theoretical comfort. The point is that it is theoretically coherent and consistent to say that roaring is used in fighting because it is energetically expensive, that nestlings beg so noisily because it reduces their growth, and that antelope stot because it reduces the speed at which they escape from lions. These apparently paradoxical ideas work because signalling systems require waste to ensure honesty. The ideas should therefore be considered along with other candidate explanations when evidence is being interpreted, and not be rejected on the grounds that they are simply absurd.

5. Conclusions from the ESS Models

The models of sections 2 and 3 are seductively general as they work with arbitrary functions w, D, G, and as they seem to have so many possible applications as discussed in section 4. In this section I consider in what ways these are not general models of signalling, and then in the light of these restrictions discuss what conclusions may nevertheless be drawn from them.

The first restriction is that the models of sections 2 and 3 are not fully general theories of information transfer. It is axiomatic in the models that the information receiver can assess perfectly the advertizing of the information provider, but cannot

assess at all the provider's quality. Just on general grounds, it is likely that both advertizing and quality are perceived imperfectly, though it is plausible that assessment of quality will be the less reliable. The second restriction is that we have assumed the information provider has perfect information about his quality, while in any real application it is likely that this will not be so.

These restrictions are not very serious, as they do not really restrict the models as models of signals. To anthropomorphize, if you observe my quality directly, that is no signal—it is the real thing. Only information "voluntarily" supplied comes under the heading of signal. Again, if I do not know my own quality, then it comes as no surprise that I cannot signal it. But in any informational transaction, the models we have seen seem so far to be reasonable models of the signal element—the element of information voluntarily provided.

The interpretation in biological terms of "voluntary" provision of information is as follows in the example of section 2. Natural selection cannot improve the quality of males, otherwise it would have done so. But natural selection can shape the advertizing level as a function of quality. So "voluntarily provided" means that natural selection could have altered the advertizing rule so that a different level was produced for the same quality of males. The interpretation of whether an individual knows his own quality is parallel. If an advertizing rule that conditions on quality can arise, ultimately by mutation, then quality is known perfectly. To the extent that advertizing can be made conditional on quality, individuals can be said to know their own quality. Consciousness is neither assumed nor excluded—it is irrelevant.

In a more realistic model, we might suppose that females obtain information both from the strategic choice of advertizing level, and in other ways over which males do not have complete control. A model of this type would be complicated because specific assumptions would be required about how information was produced by males and acquired by females, and this will not be attempted here. It is worthwhile, however, to consider a simpler model in which females obtain an estimate of a male's quality, the error of which is determined by the level of advertizing adopted by a male. High advertizing levels allow females to observe quality accurately, while low advertizing levels allow females to observe quality inaccurately.

Suppose that females observe perfectly the level of advertizing, so that they know how accurately they have assessed male quality. It is natural to guess that there will be an ESS in which good males tell the truth, as they will do well as a consequence of being accurately assessed; while poor males will advertize little to increase the chance that females will mistakenly assess them as good. This ESS does not in fact exist, and it is easy to see why not. At such an ESS, if one did exist, advertizing would correlate perfectly with male quality. Females could therefore disregard their noisy direct estimate of quality, and use advertizing levels instead. Of course once females used this rule, males would change their rule too. Females can use the information about advertizing not only for the purpose of estimating the error of their estimate of quality, but also to exploit any correlation that happens to exist for any reason between advertizing and quality. This "ESS that doesn't exist" seems to underly Zahavi's (1978) arguments on the evolution of the form of signals.

Everybody is forced to provide an equally good estimate of their quality. Note that the result depends on the perfect assessment of advertizing and the imperfect assessment of quality. If advertizing level too were assessed imperfectly, then I guess that an equilibrium could arise in which advertizing level correlated with quality, roughly in proportion to the ratio of error in assessment of advertizing level to error in assessment of quality.

Leimar (1988, chapter 2) gives models of conflict in which there is imperfect assessment of quality, and in which individuals' actions affect the degree of imperfection, but in which no inferences are drawn about an opponent's quality from his strategic choice of behaviour. This limitation on how inferences can be drawn is imposed by the model, and it would be interesting to know if Leimar's important results would be affected by allowing animals to make inferences from choice of behaviour about opponents' fighting abilities.

Returning to the limitations of the models of sections 2 and 3, they can be recognized as models of "persuasive" signalling. The male is better off the more highly females estimate his quality. The nestling is better off the more highly parents estimate his state of growth. Whether the estimate is accurate is unimportant to the signaller, indeed so long as it is in the right direction, the less accurate the better! Not all signalling is of this kind. In many cases, it matters that the estimate is accurate. This will typically be the case in co-operative endeavours, and an example might be signalling about the state of the nest between worker honey bees, If there is one singly mated queen, then there can be no conflict of interest between workers. Nevertheless, they signal to each other about the temperature of the hive, the location of food and whether the hive has been attacked. These signals would fall outside the scope of the models of sections 2 and 3. (It is probably hasty, though, to assume that if animals share the same ultimate interests then their signalling cannot contain the signs of conflict. Dr Zahavi has suggested to me that different individuals with the same interests may have different information about the world. These different states of information may be able to play somewhat the same rôle as conflicting interests, as they would lead individuals to have different expectations about the effects of the same action.)

These signals may be called "informative" signals, because their purpose is to inform, not to fool. Of course at equilibrium, as the ESS models show, even persuasive advertizing is honest and so informative; but each persuader would still like to be over-estimated, if only it did not cost him so much. Informative signals are likely to evolve to be clear with low costs: there is no waste element in them, as the interests of both parties are the same.

In many situations, signals are likely to share elements of persuasion and information. A nestling that persuaded its parent that it was too big might be mistaken for a cuckoo, or expected to fly, or be choked by feeding so fast it could not cope. The proportions of the elements may depend on the communality of interest between the interactants, and therefore on relatedness. If the queen in a honey bee nest is multiply mated, than the different patrilines of workers may have different interests. Signals about the state of the nest may then be used to try to manipulate other

patrilines to do more of the general work and defence, and less feeding of the young where there is the possibility for favouring one patriline over another in the production of young queens.

The ESS models are therefore models only of the persuasive element in signalling, in which the signaller and the receiver have different interests in the accuracy of the receiver's evaluation of the signal.

The ESS models have one formal weakness as models of persuasive signalling, which Dr Olof Leimar was kind enough to point out to me. That is that they model only the evaluative element of the response to a signal, and do not model how the receiver chooses to use the information. (I am grateful to Dr Sean Nee for this way of expressing the weakness.) The arbitrary function $w(a, p, q)$ is supposed to represent the fitness of a male that advertizes at a, is evaluated as p and has true quality q. This does not allow for the possibility that his fitness should also depend on the advertizing rule used by males, $A(q)$, or the assessment rule used by females $P(a)$. This rather abstract worry has the following instantiation. If males pay the cost of advertizing by dying, then this may reduce the rate at which females meet males; forcing them to be less choosy, and so altering the fitness consequence for a male of being perceived as of a given quality. Hence changes in $A(q)$ may bring about changes in the function w, contrary to the implicit assumption of the model. Exactly this situation arises in the population genetic model of the companion paper, which adopts a more concrete model to avoid this very problem.

The problem arises in a different way when the informational exchange is symmetrical, in the sense that both parties signal and receive information, and that the outcome breaks a symmetry to produce a "winner". Conventional fights are usually of this kind, while courtship is not. Consider a signal such as a threat display in a fight. The energetic cost is trivial, but if it is to be part of an evolutionarily stable signalling system, it must have a cost great enough to warrant the meaning ascribed to it by the receiver. That cost will be an average, taking into account cases in which the receiver flees and the cost is small, and cases in which the receiver stands his ground and makes counter threats, or actually attacks, when the cost may be very high. Now the meaning of the signal must match the average cost, using the best information available to the receiver about the best information the signaller can use to assess the cost. The cost actually paid in any one encounter depends on the strategies actually employed, and on the truth about the types of the interactants. The skill in playing the game seems to be choosing signals that are actually cheap, because of information only the signaller has, which will appear to be expensive to the receiver. But of course the receiver cannot be fooled on average, because the receiver has been selected to make the optimal choices on the information available to her. The model of section 2 can be applied to a single signal, provided the signaller and receiver are in the same informational state about the costs of the signal. But even then, the function $w(a, p, q)$ depends on the strategies employed in that particular encounter. The formal weakness therefore appears here even more radically than before.

The formal weakness can be understood in an aesthetic way as a defect in the game theory model. The female strategy set is not of definite actions that can be

taken, but consists of assessments of the male. This introduces a psychological construct into the definition of the game, a conceptually alien intrusion whose effect is to gloss over what females would actually do with the information they obtain about males. I am grateful to Professor Reinhard Selten for explaining this defect to me.

This three-pronged attack can be deflected but not rebuffed. Modelling how females use information necessarily makes the model more specific. The companion paper constructs a specific model which answers these criticisms in a particular case. I envisage that for any specific application of the model, a similarly more complete model would be required. The models of this paper are very useful as "stencils" not only for creating more specific models, but also for seeing simple ways to prove results in their more complex settings.

The game theory models can also be understood as models of models. Taking the more concrete model of the companion paper, the fitness of a male can be written down as a function of a, q, and even, by an appropriate interpretation, p. Then the results of the ESS models apply to the more concrete model. This exercise is carried out in Appendix 5 of the companion paper. As a model of a model, the results about the existence of a solution are useless, because the function w has to be taken as what is true at equilibrium in the more concrete model, and it will not in general be true that w remains unaltered when the strategies are altered. But the main handicap result of this paper's Appendix 3 still holds. The value of modelling a model concerns the interpretation of an example from nature. Supposing we have reason to believe that a signalling equilibrium exists, the main handicap result tells us that certain conditions must hold on the costs. Because no matter which more concrete model holds, and we may have little idea about what that model is like, the ESS model will hold as a model of the model; and so its conclusions must be true.

It is easier to measure the quantities in the model of the model. The fitness of different strategies could be measured by manipulating individuals, and observing how they perform in the population as it is, at its supposed equilibrium. This is a direct measure of $w(a, p, q)$. To find the fitness functions of a more concrete model, it would be necessary to know how well any strategy performed in any possible population. By concentrating on the properties of the population at equilibrium, the ESS model is therefore closer to many possible empirical tests than the more concrete models, which require more extensive information. Accordingly, the right concrete model, if it could be discovered, would provide a much fuller understanding than the ESS model of the selective forces at work in the equilibrium.

Another way in which the main models of this paper can be extended is to allow advertizing in more than one dimension. If tail length and tail breadth are used by females to assess male quality, what advertizing and assessment rules will evolve? These models are easy to write down but hard to analyse, and they suffer from the technical problem of deciding what female behaviour should be towards signals that do not appear in equilibrium. But they would nevertheless be more realistic models.

Finally, another extension is to consider that females may not all have the same interests in finding a mate. Females may have different preferences for males for

reasons of genetic compatibility, or complementarity, or differential needs for resources, say, for food and defence.

The models of sections 2 and 3 are therefore models of the evaluative element of one-dimensional persuasive signalling.

Before concluding this section, it is convenient here to consider the relationship between the models of this paper and the only previous models of biological signalling known to me, those of Enquist (1985). Enquist's purpose was to show that animals could signal intentions, contrary to claims that evolutionary game theory predicted they should not. I now describe his first model, which was a modification of the hawk–dove game, in which each player may be weak or strong, and knows initially his own strength but not his opponent's. There is first a round in which each player can choose to signal A or B, and these signals are costless and payoff-irrelevant. Then each player can decide to flee, to attack unconditionally, or to attack if the opponent does not flee. With the right combinations of payoffs, the following strategy is an ESS. Strong individuals choose A in the first round, while the weak choose B, thus making both individuals' strength common knowledge. Two strong individuals then fight each other, and two weak individuals then fight each other, in each case by unconditionally attacking; but if it turns out that one is strong and the other weak, then the weak player flees while the strong player attacks if his opponent does not flee, thus leaving the strong player in possession of the contested resource. Information is transferred by signalling in the first round.

Enquist's model differs from the ESS models of sections 2 and 3 in that it is a concrete model, with a signalling interpretation given only afterwards to the ESS. Also, it has a discrete set of possible signals. The results are nevertheless consistent with those obtained here. The sense in which his signals are costless is that there is no necessary cost to making them. Against an opponent playing the ESS, however, the consequences of signalling for the opponent's choice of action mean that they are costly in the sense used in the rest of the present paper. Furthermore, making the signal A (asserting that one is strong) is more costly for a weak individual than for a strong one. This is a good illustration of how the rather general formulation of the present paper sweeps important details under the carpet.

I now conclude the section. Despite their limitations, the models of sections 2 and 3 do provide us with some general conclusions. Persuasive signalling necessarily involves waste, as only costs enforce honesty. Further, the costs must be differential, so that it costs a better male less to make the same signal. Although the purpose of the signalling is persuasion from the signaller's point of view, the evolutionary end result is that signalling is honest and the receiver forms a correct opinion of the signaller's quality. These conclusions may be attributed to Zahavi (1975, 1977, 1987). The evolutionary stability of persuasive signalling necessitates honesty, which necessitates waste.

6. What Happened to Cheating?

The models of sections 2 and 3 show how selfish advantage is compatible with honesty. Against this background, I now consider what we think of as cheating,

beginning with a review of Zahavi's first argument, that stable signalling systems are honest. Receivers have the evolutionary option to ignore a signal, or to interpret it differently. The fact that they attend to signals and are not selected to change their interpretation of them is the sense in which signals must be honest. But the argument shows only that in some sense signals are honest "on average".

If each level of signal is made by only one type of signaller, then honesty on average implies honesty on each occasion of signalling. But suppose two types of signaller make the same signal. Then the receiver must treat those two types the same (because by hypothesis this is the only information the receiver has about the signaller). Often, one of the two types will benefit by this conflation while the other will suffer. The type that benefits can be thought of as a "cheat".

The reason that no such cheats appeared in the models in sections 2 and 3 is that the relationship between quality and marginal cost of advertizing was simple and monotonic. To produce cheats it is necessary only to suppose that as well as the original males, there is a second tranche of males that find advertizing much cheaper for a given quality of male. These males will advertize more than the first tranche, and could be considered cheats. Of course in equilibrium females would take into account what fractions of the advertizers at a particular level came from which tranche. If the extra males have too much impact on the signalling system, they will disrupt it.

The incidence of cheating must be low enough that signalling remains on average honest. As signallers maximize their fitness, this implies that the occasions on which cheating is advantageous must be limited. Perhaps the signallers for whom cheating is advantageous are in a minority, or that only on a minority of occasions does it pay a signaller to cheat. The difference between cheats and non-cheats is in the cost of the signal.

Consider Batesian mimicry as an example. The receivers are the predators. The honest signallers are the nasty bright prey, and the edible cryptic prey. The cheats are the edible bright prey and the nasty cryptic prey. The signalling must be honest on average, in that bright prey must be nastier and cryptic prey must be more edible, on average. Otherwise the predator would not respect the coloration. Now how can it be advantageous for edible prey not to be bright, when the predator is avoiding bright prey? The answer must be that brightness is disadvantageous in other aspects of the species' life, not just in its interactions with the smart predators that avoid bright prey. Perhaps other predators are not smart, or they find all the prey edible. Then the cost of the signal (being bright) is the cost induced by being bright in those other aspects of the prey species' life. Species for which those other aspects are relatively unimportant will become bright (independent of nastiness), and species for which those other aspects are relatively important will remain cryptic (independent of nastiness). The signalling system can be stable only if nasty species tend to be those for which the smart predator is a relatively important selective force, and edible species tend to be those for which the other aspects are more important. That is, if the cost of signalling is lower for the nasty species than for the edible species. The cheats comprise species which are exceptions to that general relationship between nastiness and relative importance of selective forces, namely nasty species

for which the smart predator is relatively unimportant, and edible species for which the smart predator is relatively important. Seen in this light, Batesian mimicry is one of the two types of cheating in a signalling system that falls squarely into the ambit of the model of section 2.

As a brief digression, recall that one of Zahavi's main conclusions is that the cost of a signal is a key to its meaning. A signal wasting energy shows possession of energy, a signal inviting predation demonstrates a low predation rate. This account of Batesian mimicry suggests a qualification to this conclusion. Warning coloration evolves if there happens to be a positive correlation between the nastiness of a species, and the extent to which smart predators are an important selective force in its appearance. If the correlation happens to be zero or negative, then warning coloration will not evolve. Hence the meaning of warning coloration to the predator is: "you are an important selective force in determining my appearance (because I can afford to be bright to everyone)". This indicates nastiness only because of a contingent correlation. The contingency rather spoils the usefulness of Zahavi's principle in this case.

Where there is a reason why signal cost and quality are related, Zahavi's principle applies with full force. But the example shows that it is the correlation that matters, and signals can evolve if the correlation just happens to be in the right direction. These "just happenings" imply the existence of many potential signalling systems that do not arise because for them the correlation "just happens" to be in the wrong direction. These contingent cases are more likely where the signallers are species, rather than individual types within a species, because pre-existing differences in signal costs unrelated to quality will arise there more often.

Returning to the main theme of the section, there is a second kind of behaviour which we are tempted to call cheating, often called bluffing, and is particularly likely in the symmetric informational exchanges discussed in the previous section. When a participant bluffs, she finds it advantageous to give a signal in which she benefits from the receiver's (unavoidable) conflation of the types of signaller that would have made that signal. Symmetry of the exchange makes bluffing likely, because the costs of signals are intrinsic, and a participant's best estimate of the cost of a given signal will fluctuate depending on her information about the opponent's type, and about the opponent's strategy. A signal is impressive to the receiver depending on its cost, so far as the receiver can assess it. A signal is therefore likely to be given when the signaller estimates its cost to be less than she estimates the receiver will estimate it to be. If the discrepancy between the truth about the signaller and the assessment made by the receiver is great enough, then we would call it bluff. The Zahavian condition on average honesty remains: the receiver correctly interprets the signal as having come from some probability distribution of signalling types, including the occasional bluffer, but pending further information must treat the signaller as coming from (in a loose sense) the average of that distribution.

In conclusion, cheating is expected in evolutionarily stable signalling systems, but the system can be stable only if there is some reason why on most occasions

cheating does not pay. Cheats impose a kind of tax on the meaning of the signal. The central fact about stable signalling systems is honesty, and the debasement of the meaning of the signal by cheats must be limited if stability is to be maintained.

7. What is a Signal?

This may seem a strange question to pose so late in the paper. I begin by differentiating signals from communicative manipulation, and go on to consider rather more formally what kinds of variation in behaviour should count as signals.

Signals and manipulative communication are treated synonymously by Dawkins & Krebs (1978; Krebs & Dawkins, 1984). They consider angler fish, which use a worm-like lure to attract prey. Should this lure be considered a signal? The handicap attack on understanding it as a signalling system proceeds as follows. What is the signal? It is the presence or absence of a wriggling worm-like object (that may or may not be a real wriggling worm). Who are the signallers and receivers? The signallers are real worms and angler fish. The receivers are fish that are prey to the angler fish and predators of the worms. How are the interests of the receivers served by the use of the signal? They usually find prey, and this more than compensates for the odd occasion when they are found by an angler fish instead (if it did not, they would avoid the signal, and the signalling system would collapse). So far, so good. But how are the interests of the signallers served by this signalling system? The case of the angler fish is clear, for they are a minority group parasitizing the major group of signallers, the worms. The reason worms look worm-like and wriggle is not to provide a way for predators to find them. Major selective forces on their appearance and behaviour presumably include feeding, dispersal, circulation of oxygen, and functioning of the gut. To the extent that perception by predators is important, it acts against this transfer of information. From the point of view of the worms, then, the system is not a signalling system, as their shape and behaviour are not modified in order to convey the information that they do undoubtedly convey. Our analysis, therefore, shows that the whole system is not a signalling system. The lure of an angler fish is no more a signal for fish to approach and be eaten than the firing of a shotgun is a signal that invites suicide on the part of the end users. Now Dawkins & Krebs would probably have agreed that this was true of all signals. The Zahavian approach suggests that a distinction can be drawn between on the one hand signals, in which both recipient and signaller gain by their actions, at least on average, and on the other hand manipulation, in which only one participant gains. Perhaps this is really a distinction between manipulation by use of signals, and other kinds of manipulation.

These arguments can be summarized rather more formally. Let the set of possible actions by an actor be A, and a set of character states of the actor be C. Let the set of possible actions by an observer of the action be R. Then if the choice from A is a signal for C then we must have (i) natural selection could produce any rule relating the actor's character state to his action and (ii) if natural selection can produce a rule in the observer relating A to R, then it can also produce the rule

obtained by any permutation of the elements of A. Condition (i) ensures that the signaller gains by giving the signal, and condition (ii) ensures that the actor gains by its "interpretation" of the signal. These conditions are necessary but not sufficient. They are satisfied by an action such as "when an ear itches, scratch the right ear if its the right ear that itches, and the left ear if the left itches". We have omitted any reference to the fact that the selective pressures on the actor's actions must include the response of observers. Eactly what the condition should be is unclear to me, and I now turn to an attempt to examine formally what a signal is.

The starting point will be the work of Enquist (1985), and we need to develop some concepts first. Strategies can be conditional on earlier occurrences, and specify a choice of action at each point in the game. Suppose a *subsequent action sequence* at any point in a game is a specification of a list of the actions of both players from that point of the game to the end. Then call a choice between two actions at some point in a game *payoff-irrelevant* if the payoff to each participant under any subsequent action sequence is unaffected by the choice of actions. Any effect such a choice of actions has must therefore arise only from the information provided to the opponent by the choice, and not from any strategic consequence. Of course a payoff-irrelevant choice of action used as a signal can have an effect on the payoffs—but this must be because the participants' strategies result in a different subsequent action sequence following the different choices of action.

This definition is useful for showing that any consequence of the choice of behaviours must be through the information conveyed, and not through strategic influences. It does not preclude the possibility of finding some way of teasing apart the signal element and strategic element of a choice of actions that has both. In contrast to the previous approach, Enquist's conditions are sufficient but not necessary. The whole point of handicaps as signals is that they have payoff consequences, but usually in the apparently wrong direction.

One line is to define payoff-irrelevance separately for the two players. Thus in the model of section 2, the signal of a male is payoff-irrelevant for females, but not for the male himself. Hence the advantage he gains must come from the information conveyed to females. However, even one-sided payoff-irrelevance is not true in the model of section 3, in which a male's advertizing costs are paid by the female too. Yet both cases are clearly models of signalling. One point is that their non-signal effects go the wrong way, so that their signal effects must counteract their non-signal effects before they confer any advantage at all. I am unable to offer a formal definition of signals in terms of game theory.

The problem is to tell whether a trait evolves in a model because of its strategic or signal consequences. One slightly less formal possibility is to study modifications of the game, in which the signallers are restricted to strategies that do not condition on their type, or in which receivers cannot condition on the choice of behaviour. The differences between the equilibrium states of the trait in the original model compared to the two modifications can be attributed to signal function. While for the simple kinds of models we have at the moment, definitional problems are not pressing, there is no harm in being prepared. Whether an action observed in the

field should be understood as a signal also depends on what we should want signal to mean, so the definition is not of purely theoretical interest.

8. Parallel Models in Economics

There are signalling models in economics which have strong parallels with the biological signalling models presented here. Signalling theory in economics began with Spence (1973, 1974). Riley (1979) has the closest type of model. Imagine a market with many sellers with different qualities of goods. The question is whether an equilibrium exists in which sellers of higher quality goods can advertize more and charge a higher price. Buyers are assumed to obtain information about quality only through advertizing. It turns out that, under various assumptions, such an equilibrium can exist. It is fundamentally different from the biological model in that the price is paid by the buyer to the seller. It is as if a female had to pay a male an amount related to her assessment of his quality. The level of the signal therefore affects the buyer and seller in exactly opposite ways. In the biological models developed here, if females do pay a cost of a male's advertizing, it is not a cost that benefits the male. This difference means that despite the formal similarities, the biological models and Riley's model provide little mutual enlightenment.

Cho & Kreps (1987) and Banks & Sobel (1987) discuss a methodological problem that affects signalling theory in biology as much as in economics, though the solutions may be different. In the biological model, it makes sense to say that a level of advertizing signals a male's quality. A male with a different advertizing rule will have its success evaluated in part by how females make inferences from his advertizing about his quality. If he does this by producing a level of advertizing which males of other types already make, then the females will simply erroneously conclude that he is one of that type. The problem arises if the male produces a level of advertizing that is not produced by any other males. When considering whether a candidate equilibrium really is an equilibrium, we can phrase the question as: "In calculating the success of a mutant male type, what assumption should we make about the response of females to a signal not made by any males at the (candidate) equilibrium?"

I have evaded this problem in my treatment of earlier sections, as I believe it is of merely technical interest in those models. On the other hand, in models in which male advertizing has two dimensions (say length and breadth), the problem becomes a real one. It is natural to assume that the equilibrium advertizing rule will be a path in advertizing space. Low quality males produce the combination of length and breadth at one end of the path, and higher quality males make the combinations represented by increasingly distant points of the path until the highest quality males are reached at the other end. Each type has its own unique combination of length and breadth. But to test whether a given path is an equilibrium path, we need to know how females would respond to a male choosing a combination of length and breadth not on the path.

From a purely game theoretic point of view there is no answer to this question. One natural biological response is to argue that females make only imperfect observations of advertizing levels, and so they do indeed sometimes see all possible advertizing levels. The optimal female strategy must take account of this imperfection of observation. The technical problem of equilibrium beliefs about out-of-equilibrium behaviour could be solved in some such way as this, and one of them is adopted in Appendix 1. Of course it might be biologically unrealistic to solve it in this manner. If those beliefs do drift and fluctuate because they are never or rarely tested, then this would lend a whimsical and episodic nature to signalling systems. This may be a fact about the world which should not be assumed and smoothed away.

9. Interpretations of the Handicap Principle

Zahavi's fruitful metaphor of the exaggerated sexually selected character as a test imposed on the male has been interpreted in various ways, some classified by Maynard Smith (1985). Here I discuss them, in general terms, in the light of the signalling models of sections 2 and 3. For a full review of population genetic models of the handicap principle, and reinterpretation that reconciles them to at least the sometime operation of the handicap principle, see Pomiankowski (1988).

The first interpretation of the handicap principle is that possession of the handicap imposes on its possessor extra predation risk, or extra demand for resources in the construction of the character. Only high quality males can pass the test by surviving despite these difficulties. By mating with a male that possesses the exaggerated character, a female ensures that she mates with a male who has sat an examination, and passed it by surviving. This has come to be known as Zahavi's handicap, following Maynard Smith (1985), and it does seem to be what Zahavi (1975) has in mind part of the time. The differential cost of the handicap is important, so that a higher fraction of high viability males survive than low viability males.

Models based on this first interpretation of the metaphor of a test have employed male strategies which are unconditional. A male either advertizes, whether of high or low viability, or does not advertize, whether of high or low viability. If the game theory model were restricted to such strategies, then advertizing would certainly not spread. These models rely on linkage disequilibrium to create a statistical association in the population between viability and possession of the handicap, rather than on males' flexible advertizing responses to their own viabilities. These are therefore not signalling models.

A second interpretation is that differences in quality cannot be observed by females in the ordinary way of things, but that they can be observed if males undertake some onerous task. An analogy would be that judging the physical fitness of athletes is much easier if they run a race than if they lounge around in the dressing room. This has been called the revealing handicap by Maynard Smith (1985). Zahavi seems also to have this interpretation in mind at times too. The cost of this handicap is not essential. What matters is that it is possible to disclose information in an uncheatable way—the cost of the disclosure is irrelevant. The revealing handicap does not operate as a signal, because the content of the message is directly observed.

In terms of the discussion in section 7, even artificial selection could not cause a male of low quality to appear to be a male of high quality.

The third interpretation is the condition-dependent handicap. The idea here is that only high quality males are capable of expressing the handicap. Males of low quality do not produce the handicap, and do not pay the associated costs. This way of reducing the costs associated with the propensity to have a handicap was suggested by Zahavi (1977), and later modelled by Andersson (1986). The effect of the condition-dependent handicap is very similar to the revealing handicap, except that poor quality males do not pay the cost of the handicap. The cost of the handicap is important in so far as it is responsible for the inability of low quality males to produce the handicap, which in turn is responsible for its effectiveness. Whether the condition-dependent handicap is a signal will be discussed shortly.

The fourth interpretation underlies the models presented in sections 2 and 3. I call it the "strategic choice" handicap. I believe it was in Zahavi's mind when he wrote his paper in 1975 and 1977, and it was made verbally explicit by Nur & Hasson (1984) in a paper that used mainly graphical techniques. The idea is that each male is faced with the decision of how large a handicap to incur. One of the factors relevant to this decision is his own quality, which he knows but the females do not. If the males' strategic decision is of a certain kind, males of different qualities will choose different handicap levels, thus revealing their quality to the females. The female observes the level of the handicap, infers what kind of male is in the strategic situation which would make this level optimal, and thence infers the male's quality. The work of the models is to show under what circumstances the males' situation is indeed of the type which can create this strategically induced transmission of information.

Each male can choose to produce any level of the handicap, so the strategic choice handicap is not revealing in the sense that it is constrained to be revealing. On the other hand, at equilibrium, it is true that a male's level of the handicap reveals his true quality. Similarly, the strategic choice handicap is not condition-dependent in the sense that low quality males cannot produce large handicaps. On the other hand, at equilibrium, it is true that under the males' free strategic choices a male's level of handicap is conditional on his quality. There are signs in Zahavi's papers that the strategic choice handicap is indeed what he mainly had in mind. But without a formal representation of the strategic elements of the game, it is natural to fail to state clearly the distinctions between what must be true because of physical or physiological constraints, and what turns out to be true at equilibrium as a result of optimal choices. In the many anecdotes which Zahavi retails in talks, the strategic choice handicap figures prominently.

In earlier sections we saw that for a strategic choice handicap to evolve, higher levels of the handicap must be more costly, and the marginal cost at the same level of handicap must be greater for low than for high quality males. The centrality of costs to the operation of the handicap accords with the stress Zahavi lays on costliness as a guarantee of honesty.

Returning to the question of whether condition-dependent handicaps are signals, these models do allow flexible responses to high quality males, but not to low quality

males. They work because a low quality male cannot increase the level of his handicap while the strategic choice handicap works because it would not pay him to. There is a grey area between these two interpretations. If a male could increase his level of handicap but would die very quickly, this could be seen as condition dependent, if we agreed to exclude from the strategy set those strategies which are clearly disadvantageous. On the other hand, if the level of handicap is continuous, and the effects on the male's survival are continuous, then there will be some slight increase in handicap level which does not have a drastic effect on survival, and so could not be excluded *a priori* from the strategy set. Seen in this way, the condition-dependent handicap models are approximations to the strategic choice handicap. They substitute the technically simpler device of excluding some strategies from the strategy set for the more complex alternative, representing their graded disadvantages in the model.

I therefore view the strategic handicap models as fulfilling the intentions behind the condition-dependent handicap models in a more complete way. This is in contrast with the revealing handicap, which is not a handicap at all, nor a signal; and in contrast with the so-called "Zahavi's handicap", in which males do not evolve to supply information to females either through direct observations of their quality or through signals.

10. Concluding Remarks

Some readers of an earlier version of this paper have flatteringly suggested that the signalling games are my own invention, and that the connection with Zahavi's writings on the handicap principle is rather remote. The complicated modelling, they suggest, is a long way from the Zahavi's verbal explanations. To show that the connection is strong, I want to emphasize how simple the basic arguments are. Granted that a signalling system exists, and that receivers are behaving selfishly, it must be that signalling is honest. Receivers could evolve a different rule of interpretation, but, at the equilibrium, a different rule could not be advantageous. This argument for honesty is extremely general.

Now suppose the interests of the signaller are not served by such an accurate interpretation of the signal. How can it be that the signaller does not choose to alter his signal to exploit the interpretation of the receivers? It must be that it would be costly to do so. Hence the only guarantee of honesty on the part of the signallers can be that giving what would otherwise be "advantageously untruthful" signals must be costly. Suppose further that the signallers lie on a one-dimensional continuum of the quality signalled, and that to be assessed as of higher quality is advantageous. Then for a lower quality of signaller not to gain by "pretending" to be of higher quality, it must be that the signal that means "I am of high quality" is more costly to the low quality than to the high quality male. Hence signalling more must be more costly to worse males. These two conclusions on cost apply to progressively more restricted sets of signals, but still very general sets of signals even at the end.

These verbal arguments are really just as convincing as all the mathematics, and their language makes clear the strong connection with Zahavi's arguments. This shows that the models given in this paper really are models of Zahavi's handicap principle.

The model of this paper and the companion paper are in a way models of the obvious. I certainly hope that the reader feels that the basic ideas are very simple. The need for a genetic model to show that the handicap principle could work is often expressed. There is nothing special about a genetic model. The handicap principle is a strategic principle, properly elucidated by game theory, but actually simple enough that no formal elucidation is really required.

The biologically important conclusions from these signalling models are those drawn by Zahavi (1975, 1977, 1987). The implications for sexual selection are discussed more fully in the companion paper. The handicap principle in general suggests that the form of signals may be explicable in terms of what they signal, and conversely that we may find clues about the meaning of a signal in its form. This is because the way in which the signal imposes a cost on the signaller should be directly related to what is being signalled. An animal wasting energy may well be signalling that he has plenty of energy, an animal wasting food may well be signalling that he has plenty of food. A major facet in the study of signals should be the fitness cost imposed by them. A great deal of social behaviour, including sexual selection and social dominance, can plausibly be viewed as signalling. So can much interspecific interaction, between predators and prey, dominant and subordinate species. The handicap principle lies at the heart of a whole area of biological concern.

Amotz Zahavi never gave up in trying to persuade me, along with the rest of the world, that the handicap principle was of great importance. Olof Leimar, Reinhard Selten and Sean Nee understood my models and helped me to improve them. Sean Nee and Marian Dawkins persuaded me that it was necessary to give an account of cheating, and to explain which subset of signals came under the scope of the handicap principle. Sean Nee made many helpful suggestions about presentation and content. Julee Greenough and Laurence Hurst read the manuscript and made helpful comments. I had some very useful discussions at a conference on "Evolutionary Game Theory" at the Zentrum für interdisziplinäre Forschung of the University of Bielefeld in July 1988, where in addition Roy Gardner, Bill Zame and Elinor Ostrom directed me to the relevant economics literature.

REFERENCES

ANDERSSON, M. (1986). Evolution of condition-dependent sex ornaments and mating preferences: sexual selection based on viability differences. *Evolution* **40,** 804-816.

BANKS, J. & SOBEL, J. (1987). Equilibrium selection in signaling games. *Econometrica* **55,** 647-663.

CHO, I. K. & KREPS, D. (1987). Signaling games and stable equilibria. *Q. J. Econ.* **46,** 179-221.

DAWKINS, R. & KREBS, J. R. (1978). Animal signals: information or manipulation. In: *Behavioural Ecology: An Evolutionary Approach* (Krebs, J. R. & Davies, N. B., eds) pp. 282-309. Oxford: Blackwell.

ENQUIST, M. (1985). Communication during aggressive interactions with particular reference to variation in choice of behaviour. *Animal Behav.* **33,** 1152-1161.

GRAFEN, A. (1990). Sexual selection unhandicapped by the Fisher process. *J. theor. Biol.* **144,** 475-518.

KREBS, J. R. & DAWKINS, R. (1984). Animal signals: mind-reading and manipulation. In: *Behavioural Ecology: An Evolutionary Approach* 2nd edn (Krebs, J. R. & Davies, N. B., eds) pp. 380-402. Oxford: Blackwell.

LEIMAR, O. (1988). *Evolutionary analysis of animal fighting.* Ph.D. thesis, University of Stockholm.
MAYNARD SMITH, J. (1976). Sexual selection and the handicap principle. *J. theor. Biol.* **57,** 239-242.
MAYNARD SMITH, J. (1982). *Evolution and the Theory of Games.* Cambridge: Cambridge University Press.
MAYNARD SMITH, J. (1985). Mini-review. Sexual selection, handicaps and true fitness. *J. theor. Biol.* **115,** 1-8.
NUR, N. & HASSON, O. (1984). Phenotypic plasticity and the handicap principle. *J. theor. Biol.* **110,** 275-297.
POMIANKOWSKI, A. N. (1988). The evolution of female mate preferences for male genetic quality. *Oxford Surveys evol. Biol.* **5,** 136-184.
RILEY, J. (1979). Informational equilibrium. *Econometrica* **47,** 331-359.
SPENCE, A. M. (1973). Job market signalling. *Q. J. Econ.* **90,** 225-243.
SPENCE, A. M. (1974). *Market Signaling, Information Transfer in Hiring and Related Processes.* Cambridge, MA: Harvard University Press.
ZAHAVI, A. (1975). Mate selection—a selection for a handicap. *J. theor. Biol.* **53,** 205-214.
ZAHAVI, A. (1977). The cost of honesty (Further remarks on the handicap principle). *J. theor. Biol.* **67,** 603-605.
ZAHAVI, A. (1978). Decorative patterns and the evolution of art. *New Sci.* **19,** 182-184.
ZAHAVI, A. (1987). The theory of signal selection and some of its implications. In: *International Symposium of Biological Evolution* (Delfino, V. P., ed.) Adriatica Editrice: Bari.

APPENDIX 1

Perceptual Inaccuracy and Local Flat Extrapolation

In signalling games such as those analysed in this paper, there often exists one or a few "reasonable" equilibria and very many equilibria that seem "unreasonable". There is a whole literature in game theory concerned with establishing formal criteria for picking out the "reasonable" equilibria. The reason for this is discussed in section 8. It is that the stability of a candidate equilibrium can depend on beliefs about the meaning of signals that are never actually given at the candidate equilibrium. Beliefs about signals that are sometimes given are constrained to fit the facts, as untrue beliefs lead to loss of payoff when they are put to the test. But there is nothing in the definition of equilibrium to constrain beliefs about signals not given at the candidate equilibrium. It is beyond the scope of this paper to offer a general solution to this large problem, but it is useful to have a result that in effect says we have discovered all the "reasonable" ESSs in a game.

The basis for our criterion will be perceptual inaccuracies on the part of females (to use the example of mate choice again). The idea is that a female makes slight errors in assessing the level of a signal. Suppose signals are given in the range $[a_1, a_2]$ but not in the range (a_2, a_3). Then females sometimes observe signals just above a_2, as the result of misperceiving a signal just below a_2. The "evolutionary experience" of females is therefore that males just above a_2 are of about the same quality as males just below a_2. The out of equilibrium behaviour of advertizing at just above a_2 will therefore be met by an evolved, constrained response, and not by the arbitrary response that would result if females never perceived those advertizing levels.

I shall therefore make the following assumption. If a signal not given in candidate equilibrium is sufficiently close to a signal that is given, then females will interpret the signal as if it had been the closest given signal. This is local flat extrapolation. Local because I assume nothing about signals far from given signals. Flat because

females do not extrapolate using the slope of advertizing on quality at the nearest given signal. This formal criterion involves a leap of logic from its motivation. A fuller method would take into account explicitly, in the model itself, the errors of perception, rather than encapsulating them into an add-on formal criterion for selecting equilibria. But this paper is not chiefly concerned with this technical problem. Note I have assumed that the set of given signals is closed, so that a "closest" signal exists. A more general formulation is given in Appendix 4 of the companion paper.

There is one application of this criterion here, and more in the companion paper. We establish here that the minimum quality male must advertize at the minimum advertizing level. Suppose not. Then females would treat males with a level of advertizing just below the lowest given in the same way as they treat males giving the lowest level. Males giving the lowest level can therefore gain by reducing their advertizing. They save on advertizing costs, and suffer no loss in mating success. This supplies a meaning for the minimum advertizing level, as the level below which the marginal cost of advertizing is negative. Too short a tail would hinder a peacock in balance and flight.

APPENDIX 2

The Basic ESS Model

The existence of a solution to the ESS conditions (1) is to be proved, and that it is of the form given in (2).

Inspection of the first order condition for A suggests defining the function P^* over the interval $[a_{\min}, \infty)$ by:

$$P^*(a_{\min}) = q_{\min}$$

$$P^{*\prime}(a) = -\frac{w_1[a, P^*(a), P^*(a)]}{w_2[a, P^*(a), P^*(a)]}.$$

This unambiguously defines P^* over the whole interval, in view of the assumption that w_1/w_2 exists. Note that P^* is monotone increasing because w_1 and w_2 are of opposite sign by assumption. This allows us to define A^* by

$$P^*[A^*(q)] = q \quad \forall q \in [q_{\min}, q_{\max}].$$

I now proceed to prove that (A^*, P^*) is an evolutionarily stable pair of strategies.

The marginal value of advertizing, given that females use the rule P^*, is

$$\frac{\partial}{\partial a} w[a, P^*(a), q] = w_1[a, P^*(a), q] + P^{*\prime}(a) w_2[a, P^*(a), q].$$

When we substitute for $P^{*\prime}$ using the defining eqn in (2), and divide by $w_2[a, P^*(a), q]$, which is positive by assumption, we obtain that the marginal value of advertizing has the same sign as

$$\frac{w_1[a, P^*(a), q]}{w_2[a, P^*(a), q]} - \frac{w_1[a, P^*(a), P^*(a)]}{w_2[a, P^*(a), P^*(a)]}.$$

Now consider the left hand quotient. If this is an increasing function of q, then the marginal value of advertizing is negative for $q < P^*(a)$, zero for $q = P^*(a)$, and positive for $q > P^*(a)$. Since $P^*(a)$ is an increasing function with inverse A^*, this implies that the marginal value of advertizing is positive if $a < A^*(q)$, zero if $a = A^*(q)$, and negative if $a > A^*(q)$. Hence if the left hand quotient is an increasing function of q, $A^*(q)$ is a globally best strategy. By a similar argument, if the left hand quotient is increasing in q in the neighbourhood of $a = A^*(q)$, then A^* is a locally best strategy. Conversely, if A^* is a strictly best strategy locally, then the left hand quotient is increasing in q at least locally.

In applications, it will often be easy to show that

$$\frac{w_1(a, p, q)}{w_2(a, p, q)}$$

is strictly increasing in q for all a, p, and that therefore whatever function $P^*(a)$ is defined by (2) would be stable.

It has now been established that the ESS condition on A^* is satisfied. The ESS condition on P^* is easily seen to be satisfied as well. The integral in the minimand is of non-negative terms, so its minimum possible value is zero. When $P^*[A^*(q)] = q$, as it is by construction, this minimum is achieved. The pair of functions (A^*, P^*) is therefore an ESS as claimed.

Maximization of w depends just on the ordering of values of w and its arguments. In accordance with this, all the conditions on w would be unchanged for a function $\hat{w}$ defined by

$$\hat{w}(a, p, t) = u_w[w(u_a(a), u_p(p), u_t(t))],$$

where the functions u are arbitrary increasing invertible differentiable functions.

APPENDIX 3

The Backward Result

The result to be proved is:
If A, P is an ESS, $w_2 > 0$ and $A(q)$ is increasing, then

(a) $P[A(q)] = q \quad \forall q$ (honesty)

(b) $w_1 < 0$ (cost)

(c) $\dfrac{w_1(a, p, q)}{w_2(a, p, q)}$ is strictly increasing in q near the path $[A(q), q, q]$ (costlier for worse males).

The fact that $A(q)$ is increasing means it has an inverse function A^{-1} mapping the range of advertizing levels played for any quality into the possible qualities. This function when employed by females ensures that D is minimized. Hence $P = A^{-1}$, proving part (a). The first order condition for $A(q)$ to maximize $w(a, p, q)$ is

$$w_1 + P' w_2 = 0.$$

But P' is positive because A' is, and w_2 is positive too, so w_1 must be negative, proving part (b). If A, P is an ESS then A is a global and therefore a local maximum. For the first order condition to yield a local maximum rather than a minimum, we require part (c), as seen in Appendix 2. This completes the proof.

APPENDIX 4

The "Females Pay Costs" ESS Model

It is to be shown that A^*, P^* defined in (4) is a solution to the ESS conditions (3). The first order condition for the male maximization is

$$\frac{\partial}{\partial a} w[P^*(a), v(a, q)] = P^{*\prime}(a) w_1[P^*(a), v(a, q)] + v_1(a, q) w_2[P^*(a), v(a, q)] = 0.$$

Substituting $Q(a)$ for q defined by $Q[A^*(q)] = q$, we obtain

$$P^{*\prime}(a) w_1[P^*(a), Q(a)] + v_1[a, Q(a)] w_2[P^*(a), v(a, Q(a))] = 0$$

$$\forall a: Q(a) \in [q_{\min}, q_{\max}].$$

Now substituting for $P^{*\prime}$ in terms of Q', using $v[a, Q(a)] = P^*(a)$, provides us with the differential equation for $Q(a)$ given in (4). $Q(a_{\min}) = q_{\min}$ according to Appendix 1. It is easily verified that the solutions for A^*, P^* given in (4) solve this first order equation for males.

To tackle the second order condition, we follow the method of Appendix 2, by substituting into the marginal value of advertizing using the formula for $P^{*\prime}$. After dividing by $w_1[P^*(a), v(a, q)]$, which is positive, we see that the marginal value of advertizing has the same sign as

$$\frac{v_1(a, q) w_2[P^*(a), v(a, q)]}{w_1[P^*(a), v(a, q)]} - \frac{v_1[a, Q(a)] w_2[P^*(a), v(a, Q(a))]}{w_1[P^*(a), v(a, Q(a))]}.$$

If the left hand quotient is an increasing function of q, then the marginal value of advertizing is negative for $q < Q(a)$, and positive for $q > Q(a)$. As $Q(a)$ is increasing, the inverse function A^* is also increasing. This implies that the marginal value of advertizing is positive for $a < A^*(q)$, and negative for $a > A^*(q)$. If the left hand quotient is globally increasing in q, then $A^*(q)$ is a globally best strategy, while if it is increasing only local to $q = Q(a)$, then we can conclude only that $A^*(q)$ is locally best. Hence the $A(q)$ is globally stable if

$$\frac{v_1(a, q) w_2[P^*(a), v(a, q)]}{w_1[P^*(a), q]}$$

is increasing in q. Strict local stability holds if it is strictly increasing in q near the path $a = A(q)$.

As in Appendix 2 the stability concerns the monotonicity with respect to true quality of (minus) the expression for $P^{*\prime}$ with general arguments. Again its

interpretation is that males of higher quality must suffer less from the deleterious effects of advertizing (that effect is $v_1 w_2$, the numerator) than their capacity to gain from extra attractiveness suffers (w_1, the denominator).

The female minimization is easily seen to be satisfied because their assessment is always correct so the arguments of D are equal. The non-negative integral therefore equals zero and so achieves a minimum. It is a strict minimum because any other choice for $P(a)$ which differs on a set of positive measure will contribute positive terms to the integral. A^*, P^* is therefore an ESS pair of functions as claimed.