ADVANCES IN THE STUDY OF BEHAVIOR, VOL. 26

How To Avoid Seven Deadly Sins in the Study of Behavior

MANFRED MILINSKI

ABTEILUNG VERHALTENSÖKOLOGIE
ZOOLOGISCHES INSTITUT
UNIVERSITÄT BERN
HINTERKAPPELEN, SWITZERLAND

Scientific journals have become both thicker and more numerous during the last twenty years. An enormous flood of papers has driven the traditional refereeing system to its limits and scientists to increasing specialization. For someone of my generation who originally analyzed his data with pencil and paper, drew his figures with ink, and wrote his first publications on a mechanical typewriter, producing papers in our software age has become extremely easy and many times faster. Two steps in the paper milling process must still consume much time and effort, namely designing and performing experiments. The consequence of speeding these steps up is a high rate of flaws and mistakes in papers, even those that are published in respected journals; obviously, referees cannot manage (or are not trained) to find all the procedural mistakes when they do their altruistic job in their limited time. I should not be too pessimistic. I am happy to acknowledge that the application of statistical procedures has been improved on average in behavior research and that there are exceptional scientists who do high quality research despite a high publication rate.

The following is a list of mistakes that I have often found in published research:

1. Unjustified conclusions are made from observational (i.e., correlational) data.
2. Data are not independent ("pseudoreplication").
3. Treatments are confounded by time and sequence effects.
4. No efforts are made to avoid observer bias.
5. Potential artifacts arise when animals are not accustomed to experimental procedures.

6. Unsuitable controls are used.

7. An attempt is made to "prove" the null hypothesis with small samples.

This list is not complete but I think future publications would be improved to a large extent if these mistakes were avoided. There are excellent reviews on experimental design (e.g., Hurlbert, 1984; Martin and Bateson, 1986, 1993) that also discuss these mistakes and give advice on how to avoid them. Why has this advice been ignored so often? Reading a very detailed review or book might be regarded as too time consuming (although performing a flawed experiment is a complete waste of time); the advice might be too complex to be easily remembered; the mistakes might be spotted but they are difficult and/or expensive to avoid ("making the best of a bad job"); the mistakes might be regarded as "accepted" because they are often made; "habituation" of both researchers and referees might be occurring.

It is worthwhile to write another advice paper only if this advice is more likely to be accepted and used. To achieve this I concentrate on only a few, but important, mistakes. I avoid giving detailed information and present only a few theoretical or technical explanations. Instead, I concentrate on simple but telling examples and incorrect and correct conclusions. I hope to be successful with this procedure because empiricists such as myself can handle examples better than we can handle theory, and because our intuition is trained by understanding examples. I do not cite specific negative examples from the literature, not even from my own research, although I have to admit that I have made one or another of these mistakes myself. Instead, I invent simple cases or cite positive examples and justified conclusions from the literature, and particularly from my own work because I know its weaknesses best and can discuss them. One of my colleagues is convinced that there is no such thing as a perfect experiment: let us try to find a close approximation to it.

I. Unjustified Conclusions from Observational Data

To observe animals and to draw conclusions about the function and the mechanisms of the observed behavior has a long tradition, particularly in ethology. What is wrong with this approach? It would be acceptable if observed correlations were only described and no cause-effect relationship conclusion was drawn from them. Let us take three examples.

A. Correct Description

1. Conspicuous individuals have a higher predation risk.

2. Smokers have a higher risk of lung cancer.

3. Ornamented males are more attractive to females.

B. UNJUSTIFIED CONCLUSION

1. Conspicuous individuals have a higher predation risk because they are conspicuous.

2. Smokers have a higher risk of lung cancer because of smoking.

3. Ornamented males are more attractive to females because they are ornamented.

C. WHY ARE THESE CONCLUSIONS UNJUSTIFIED?

1. Conspicuous individuals are, for example, larger (or have some other trait by which they differ from cryptic individuals); they may have a higher predation risk because they are larger and not because they are conspicuous.

2. People who smoke may do so because they have, for example, a gene that makes them like tobacco; the same gene may predispose them to lung cancer; they may develop lung cancer because they have that gene and not because they smoke. This hypothetical example demonstrates the weakness of epidemiological evidence.

'3. Ornamented males may be stronger, and females may like these males because they are stronger and not because they are ornamented.

D. DO AN EXPERIMENT

Example 1: What can be done if one wants to determine whether an observed trait A (e.g., high predation risk) is caused by an observed trait B (e.g., being conspicuous)? One must do an experiment, in which one removes the possibility that a difference in predation risk between conspicuous and cryptic individuals is caused by any other (undetected) trait that conspicuous individuals usually possess (e.g., larger size). One determines randomly which of, for example, 48 experimental animals has to be made conspicuous and which cryptic.

Procedure: Take two animals, toss a coin to determine which one is to be made conspicuous. Take another two animals, toss a coin, and so on, until there are 24 animals that are to be made conspicuous and 24 that are to be made cryptic. Present a predator with a simultaneous choice between an individual that has been made conspicuous and an individual that has been made cryptic. Repeat this procedure with each pair separately (independently). If conspicuous individuals are preferred over cryptic individuals (sign test), one may conclude that the specific conspicuousness that has been tested increases an individual's risk of predation. I should, however, emphasize that a statistical test provides only a nonzero probability, and

that one can be only reasonably confident in a result, but never certain (many of today's computer programs round off p at the third decimal place, tempting users to uncorrectly report a p equal to zero). By this procedure, all other traits that may affect an individual's predation risk (e.g., size) were assigned randomly to the experimental (conspicuous) and control (cryptic) groups. Randomization is therefore indispensable. We assume that the experiment has been performed properly in all other respects. There are, however, other possibilities for making mistakes that I will discuss later. I will use this example throughout this review. It turns out that this apparently simple experiment is extremely difficult to do properly.

E. An Experiment May Be Unethical

Example 2: To determine experimentally whether smoking causes lung cancer one would have to assign young adult people randomly (indispensable condition!) to the group who has to smoke, say, for the next twenty years and the group who must refrain from smoking for the same period. If the smokers have developed lung cancer significantly more often than the nonsmokers, one may conclude that smoking causes lung cancer. Of course, this experiment cannot be done for ethical reasons. What can be done instead? Should one believe in the correlational (epidemiological) evidence because it is the best we have? I do not smoke because I "believe" that smoking causes lung cancer. However, nobody is forced to believe such a conclusion. Correlational evidence can never prove a cause-effect relationship. Other problems may lead an experiment to be considered either unethical, or too difficult or too expensive to do. All this is no excuse for accepting correlational evidence as proof of a cause-effect relationship.

F. Another Kind of Experiment

Example 3: One can do the same kind of experiment as proposed for Example 1 to test whether females prefer ornamented males because they are ornamented: males are selected randomly and supplied with either a small or a large ornament. This has been done successfully by elongating or shortening the tail feathers of widowbirds (*Euplectes progne*) by Andersson (1982) and of barn swallows (*Hirundo rustica*) by Møller (1988). It may, however, be difficult to supply male sticklebacks *(Gasterosteus aculeatus)* with a red belly without producing artifacts. The solution would be to establish that females prefer males that are naturally redder. This is a staged observation (not an experiment) and provides only correlational evidence. The next step would be to prevent the females from seeing the natural differences in male red coloration by repeating the previous staged observa-

tion with other females under filtered (green) light (Milinski and Bakker, 1990), where differences in red cannot be detected. If one finds a significant decrease of the earlier preference, one may conclude that females prefer males more when they are redder. This is experimental evidence. The problem is that one has to be lucky that females do not use other male traits that correlate with red color and that are still detectable under green light. We were lucky.

G. Three Further Examples to Train One's Intuition

Example 4: If someone does not know the price of either a Rolls Royce or a Volkswagen Rabbit and looks instead at what is left in the bank accounts of people who just bought either a Rolls Royce or a Rabbit, the person would conclude that a Rabbit is much more expensive than a Rolls. The person is mistaken because rich people can buy expensive things and still have more remaining. The correlational evidence is misleading. The necessary experiment would be to force randomly chosen people to buy either a Rolls Royce or a Rabbit. I bet that only the new Rabbit owners would have on average a positive bank account after the purchase.

Example 5: Male secondary sexual ornaments are regarded as handicaps for survival. Møller (1990a) observed that male barn swallows that had naturally longer tail feathers arrived earlier at their breeding sites (from their overwintering sites) than short-tailed males. Does this correlational evidence mean that elongated tails improve flying rather than handicap it? No, it needs an experiment for such a conclusion. Møller (1989) found that the males with experimentally elongated tail feathers (compared to males with experimentally shortened tail feathers) had viability costs, such as impaired foraging efficiency, which proved that this ornament was indeed a handicap. Why did males with naturally longer tails have an improved flying ability? As natural Rolls Royce buyers, male barn swallows that grew a longer tail had more reserves than others (Møller, 1989). Stronger males should invest an amount of reserves in larger ornaments so that they are, nevertheless, still stronger than weaker males that invested in the smaller ornament (Grafen, 1990).

Example 6: Small passerine birds that have detected a predator, for example, an owl or a kestrel, approach it closely at which time they call and flick their wings. The function of this "mobbing" behavior may be the tradition of knowledge of predators passed on to offspring (Curio, Ernst, and Vieth, 1978) and/or making the predator "move on" to another site where it has not yet been detected (Pettifor, 1990). Is approaching the predator risky? It has almost never been observed that a mobbing bird is attacked by the predator that is being mobbed. However, it may well be

that the birds take their own fleeing ability into account when they decide on their mobbing distance. Curio (1983) found that birds of species that have a high maneuverability have a closer minimum mobbing distance than birds of species that can fly less well. Birds that approach a predator more closely may less often be taken by the predator than birds that approach the predator less closely (analogous to Rolls Royce buyers and naturally long-tailed male barn swallows). To test whether predation risk increases with shorter distance to the predator, one has to assign randomly chosen individuals to different approach distances and measure their probability of being caught. This has been done with dead sticklebacks that we made by remote control to "inspect" a predatory pike (*Esox lucius*) up to predetermined distances; analogous to experimentally elongated barn swallows, experimentally inspecting sticklebacks have increasing costs when they approach a pike more closely: they are more likely to be caught.

H. Is Correlational Evidence of Any Value?

Yes, correlational evidence is useful whenever it is important to know a relationship between traits. We expect both richer people and richer barn swallows to invest more in handicaps and remain still richer than poorer individuals with smaller handicaps. To establish this correlation provides some support for the ESS prediction. Knowing that predator approach distance correlates negatively with a mobbing bird's fleeing ability helps to predict differential mortality in a given situation. We regard the correlation as a matter of fact but refrain from concluding any cause-effect relationship. Furthermore, correlational evidence may be enough if a visible trait can be used as an indicator of another trait that is invisible. If female barn swallows benefit from mating with males that are genetically resistant to mites, the females may find these males by selecting long-tailed ones, because the mite load of a male's offspring correlates negatively with their father's tail length (Møller, 1990b). Similarly, the intensity of male sticklebacks' red breeding coloration correlates positively with their physical condition (Milinski and Bakker, 1990). Females can use the existence of this correlation for selecting males of superior condition.

It is not easy to find more applications for using correlational evidence correctly. Another example might be that an interesting correlation that is found unexpectedly might help to generate hypotheses for a decisive experiment; for example, Wedekind, Seebeck, Bettens, and Paepke, (1995) found that women prefer the smell of men to whom they are dissimilar in their MHC-alleles, which may be adaptive because the offspring would resist a broad range of infectious diseases with their large number of different MHC-alleles; however, women who take the contraceptive pill prefer

the smell of men with similar MHC genetics. Because this is only correlational evidence, people would like to know whether the pill actually causes this switch of preference. If so, there should be a warning on the package.

It is sometimes suggested that if the direction (e.g., positive) of a correlation is predicted under the assumption of a specific cause-effect relationship, then this cause-effect relationship is supported when that correlation is found. This is mistaken. Imagine the following scenario: some hypothetical biologists are asked to investigate whether animals that live close to motorways are negatively affected by the traffic. If anything, one expects a reduction of condition near the motorway. The biologists measure the reproductive success of rabbits that live either close to a motorway or a hundred meters away from it. They find that rabbits that breed next to the motorway have indeed significantly reduced reproductive success. They erronously conclude that motorways cause a reduction of reproductive success in rabbits. Why are they mistaken? Imagine that rabbits do not prefer territories next to the motorway (perhaps because foxes prefer those places for some reason). This has the consequence that the stronger rabbits win the preferred places farther away from the motorway and the weaker rabbits must settle close to the traffic. Weaker rabbits may have a lower reproductive success anyway, and this continues when they settle close to the motorway. Our biologists would have been quite surprised if they had found that rabbits close to the motorway have an increased reproductive success (e.g., if foxes hated noise and therefore the stronger rabbits fight successfully for places near the traffic). The only way to solve the problem is to do a suitable experiment.

II. Data Are Not Independent: "Pseudoreplication"

The term pseudoreplication was proposed by Hurlbert (1984). He found that pseudoreplication occurred in almost 50% of publications in respected ecological journals. People have become aware of this problem in the meantime and Hurlbert's paper has become a "citation classic" (see also Kroodsma, 1986, 1989). However, pseudoreplication is still found in many recent publications, perhaps because it is sometimes difficult to detect; "it can appear in different guises" (Hurlbert, 1984). Pseudoreplication occurs when replicates are not statistically independent.

A. Pseudoreplication in Experiments

When the experiment I proposed for Example 1 is performed, there are several opportunities for pseudoreplication. When pairs of fish are taken

and a coin is tossed to decide which fish has to be made conspicuous and which cryptic, both sorts of fish must be housed until the predation experiment. We have two separate tanks, one for the conspicuous fish and one for the cryptic fish. Of course (hopefully) the tanks are the same size, have the same water level, light, equipment, and so on. Unfortunately, we are not aware that something happened to one of the tanks before the experiment (e.g., someone unwittingly hit the table with the tank that contained the fish that will be made cryptic). Because of this event all the cryptic fish have been frightened and will, therefore, be more cautious in the predation experiment and thus less frequently preyed upon than the conspicuous fish. We conclude erronously that a prey's conspicuousness increases its risk of predation. This kind of pseudoreplication can be avoided by maintaining each fish in a separate tank. Of course tanks of conspicuous fish have to be interspersed with tanks of cryptic fish, otherwise hitting one table would affect one group more than the other.

One wants to investigate whether conspicuousness increases a prey's risk of predation. Another opportunity for pseudoreplication: 24 pairs each consisting of a conspicuous and a cryptic fish and 6 pike are available. Each pike is tested with four pairs of fish, one pair per day on four consecutive days. How large is n, 24 or 6? Whether conspicuousness increases a prey's risk of predation depends on the preference of the predator. Therefore, actually the behaviour of the pike is being investigated, and each pike has its individuality. The four choices of each pike are not independent because they are made by the same pike. To avoid pseudoreplication, each pike has to be treated as a statistical unit and the choices of each pike must be entered as one data point in the analysis (e.g., percentage of conspicuous fish chosen). Thus, n is only 6; an n of 24 would be pseudoreplication. What if there is only one pair of a conspicuous and a cryptic fish that are confined in bottles and presented to all pike sequentially? Our n is 1 in this case, and anything else would be pseudoreplication (the cryptic fish could have been frightened, more satiated, etc.).

One still wants to investigate whether conspicuousness increases a prey's risk of predation. What if 24 pairs of prey fish are available but only 1 pike, named Fisher. This case is not as bad as the previous one. One will find out a lot about this individual pike. One cannot, however, conclude that pike prefer conspicuous prey (or even that predators prefer conspicuous prey). One may conclude that Fisher prefers conspicuous prey, which does not help much to understand pike-prey fish relationships in general; perhaps Fisher prefers conspicuous prey because of a terrifying experience he had with a cryptic bait from a fisherman. Results from a single animal can be very valuable when it is tested for its abilities. If one chimpanzee can be taught a sign language that she uses afterward to communicate in a

sensible way with people (Gardner and Gardner, 1969), one knows what chimpanzees could do in principle. Similarly, I would be impressed if a single fish would solve a very complex optimal foraging task-at the risk that I have tested the R. A. Fisher among all fish by chance.

B. Pseudoreplication in Correlational Studies: Nessy and Loch Nessyless

It is common knowledge that "Nessy" lives in Loch Ness and it is known for sure that there is no such thing in Loch Nessyless 50 km away. One wants to know whether the fish in Loch Ness have developed some sort of anti-Nessy behavior. If this is the case, this knowledge would be of enormous importance for tourism because one can test the fish of all the other 200 lochs and tell whether they contain a Nessy. Fish of the same species, sex, size, and age are collected from both Loch Ness and Loch Nessyless and maintained in individual interspersed tanks until testing with a sophisticated model of Nessy. One does not need to test lab-bred offspring of those fish because it does not matter whether a Nessy recognition is learned or genetic. In the test the fright reaction of individual fish is quantified with the result that fish from Loch Ness are more readily frightened by a dummy Nessy than fish from Loch Nessyless. What can be concluded? The problem is similar to that of maintaining conspicuous fish in one tank and cryptic fish in another tank (see previous discussion). Many things might have happened in Loch Ness that did not happen in Loch Nessyless, which make fish from Loch Ness more cautious than fish from Loch Nessyless. It would be pseudoreplication to compare fish from only two populations that differ both in the trait of interest and in many other traits that one is not aware of.

Would it help if six different lochs that each have a Nessy can be compared with six other lochs without a Nessy? Yes, this would be much better. Now, a loch would be the statistical unit and n is 12. Pseudoreplication has been avoided, but the results would still suffer from being only correlational evidence. The lochs that have a Nessy were probably more suitable for Nessies than the lochs without a Nessy. That means that Nessy lochs have a number of traits in common by which they differ from Nessyless lochs. One can hardly rule out that one or several of these traits are responsible for the more cautious behavior of the fish from Nessy lochs and not the presence of Nessies themselves.

Are population comparisons completely useless? No, there is a way to improve them. Suppose the dummy Nessy has less pronounced effects on the cautious behavior of the fish of Nessy lochs when the dummy is made step-by-step less realistic. One, therefore, has the impression that the com-

plex Gestalt of Nessy matters for the fish. Moreover, if the measures of anti-Nessy behavior include specific behaviors that are suitable to avoid Nessy but not other large objects, then even sceptics might accept the possibility that anti-Nessy behavior has been developed by fish from Nessy Lochs. There are some examples of the latter kind of population comparison. Magurran (1990a) found that offspring of minnows (*Phoxinus phoxinus*) from a population with pike developed a sophisticated antipike behavior when presented with a realistic dummy pike; this behavior was much less pronounced and sophisticated in minnow offspring from pike-free waters.

III. Treatments Are Confounded by Time and Sequence Effects

A. Selection of Subjects

Imagine that we had not caught pairs of fish from our storage tank and had not decided, by tossing a coin, which should become conspicuous and which cryptic of each pair. Instead, the 24 fish that were to become conspicuous were caught first, and second were caught the 24 fish that were to become cryptic. By this procedure the two experimental groups would be biased by two different sequence effects: the first fish that are caught would be the easiest to catch, that is, the least cautious fish, whereas the last fish would be the most difficult to catch, that is, the most cautious fish. Consistent temperament differences, shy versus bold, among individual fish have been found in many species of fish (e.g., Huntingford, 1976; Milinski, 1987; Magurran, 1993) and other animal species including man (Wilson, Clark, Coleman, and Dearstyne, 1994). A second sequence effect would inevitably come about by disturbing the fish each time one is caught. Fish that are caught early in the sequence are less often disturbed than fish that are caught late in the sequence. Both sequence effects combine to place the more cautious and the more often disturbed fish in the group that are made cryptic and the less cautious and less often disturbed fish in the group that are made conspicuous. It would thus come as no surprise if the conspicuous fish were preyed upon more frequently by the pike, not necessarily because they were conspicuous but possibly because they were the less wary fish because of sequence effects. The elegant solution to this problem is to catch two fish (even sequentially) and determine by tossing a coin which becomes a treatment fish and which a control fish. In this way all time and sequence effects will affect treatment and control groups similarly.

B. Sequence of Treatment and Control

Again, we already have the best sequence of treatment and control subjects by presenting both the conspicuous fish and the cryptic fish simulta-

neously to a pike; one has to arrange only that the side on which the conspicuous fish is presented is randomized between trials. Now, assume that a simultaneous choice situation is not possible for technical reasons and one has to present each prey fish singly in a sequential choice situation and measure the latency of the pike's attack. If one started with all the conspicuous fish followed by all the cryptic fish for each pike, one's results would again suffer from time and sequence effects: the pike may become faster (or slower) during their time in the experiment because they learn about the prey, become more accustomed to the procedure, become hungrier, simply become older, and so on. The results would be of little value. How should one proceed instead? R. A. Fisher would have suggested strict randomization of the sequence of treatment and control (see Hurlbert, 1984). I did this when I performed my first experiment for my Master's Thesis 22 years ago and obtained the following sequence of eight trials: T, T, T, C, T, C, C, C. Although I had randomized, this was almost no better than having all treatments before all controls. The best solution is probably, again, to have pairs of treatment and control and determine each time by tossing a coin which is first and which is second. The way in which animals were assigned should be described in detail in the Methods section so that readers can judge for themselves whether the method that was used was suitable to achieve randomization. One should not say "animals were assigned randomly to treatment and control," but say "animals were assigned to groups applying the following constraints: within each pair selected a random choice was made of which would be the treatment and which the control animal."

IV. No Efforts To Avoid Observer Bias

I do not assume that you and I want to manipulate results deliberately. It is, however, known from psychological studies that people tend to interpret unconsciously what they see in a way that fits their expectations better. This phenomenon is called observer bias (e.g., Martin and Bateson, 1993). How does it occur? In the early days of ethology, researchers often just observed a subject's behavior and described what they had seen. This method allows observer bias, as can be shown with Example 1. Conspicuous prey may have a higher predation risk because they were more easily detected than cryptic prey. In Experiment 1 this hypothesis can be tested by determining the prey type that the pike approaches first. Approach means "moving toward something." There are two opportunities for interpretation: the length of a move can vary between long (e.g., 20 cm; see Fig. 1A) and just recognizable (a few mm; see Fib. 1B). A just recognizable move might be interpreted as a "move" when toward the conspicuous prey

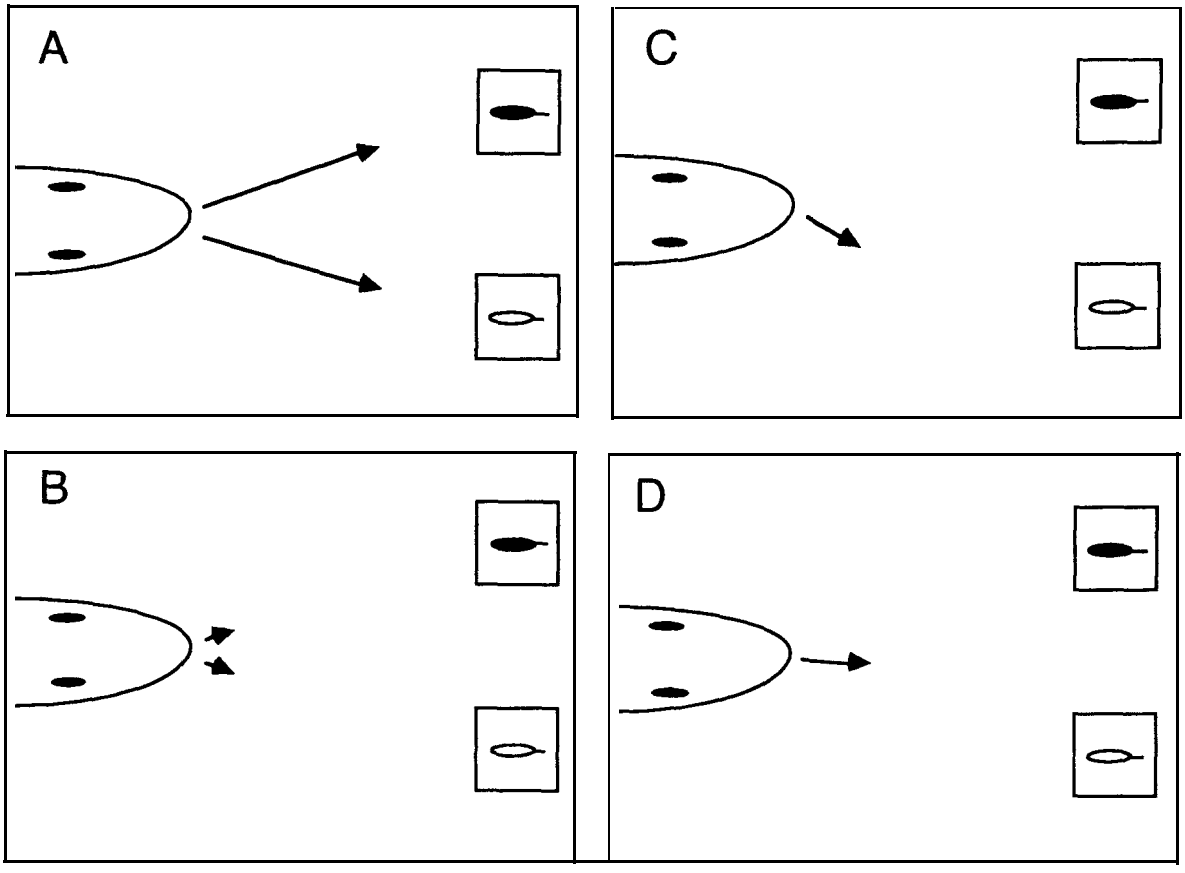

FIG. 1. A pike is observed to approach either the conspicuous or the cryptic prey. (A) A move over a long distance will always be regarded as a "move." (B) A move over a very short distance may be regarded as a "move" only when it is directed toward the conspicuous prey. (C) The direction of a move may be noted correctly when it is on a straight line toward a prey. (D) When the direction of a move is between the conspicuous and the cryptic prey, one may be inclined to regard it as being directed toward the conspicuous prey (or what is expected).

and as "no move" when toward the cryptic prey. "Toward" may vary between "in a straight line to the cryptic prey" (Fig. 1C) and "more toward the cryptic prey than toward the conspicuous prey" (Fig. 1D). In the latter case one might be inclined to see the direction of a move as being toward the conspicuous prey when it is actually toward the cryptic prey. How can we solve this problem?

A. UNBIASED OBSERVATIONS IN THE FIELD

One could ask someone to do the observation who does not know one's hypothesis. The problem here is that because the two prey types are obviously different, the naive observer can unconsciously invent a hypothesis about the type of prey that the pike should prefer and bias his observations accordingly. Another problem is that the observer may miss detecting the

cryptic prey more often than the conspicuous prey. So this does not work. Instead, one could make a video recording that is shown to naive persons so that the difference in conspicuousness between the two prey types is not detectable (e.g., color turned off, if the difference is in color). With this procedure observer bias is avoided, but the results will suffer from a large variance because the observer may often make mistakes in detecting just recognizable approaches.

B. Unbiased Observations under Experimental Conditions

In an experiment it is much easier, but technically demanding, to avoid observer bias by defining exactly "approach toward." A naive observer is not needed. We use again Example 1. A tank is divided into halves by an opaque partition that has a sliding door in the middle. The two prey fish are each confined in a Plexiglas container in one side of the tank at the same distance vis-a-vis the sliding door. The pike lives in the other half of the tank. One such tank is needed for each pike. When the sliding door is lifted, by remote control, at a preset time (otherwise one may be inclined to lift it when the conspicuous fish is moving!), a video camera records the pike's behavior from above. Two thin lines are drawn on the video screen: one parallel to the opaque partition at a distance from the door where one finds it useful to determine the pike's decision, another line that cuts that half of the tank into a left and a right part as seen by a pike that enters by the door (Fig. 2). Now, one has almost no opportunity for subjective bias. One determines whether the tip of the pike's snout is in the right or the left part when it passes the decision line. Unfortunately, this rather complicated design is necessary for obtaining unbiased observations.

C. When Behavior Is Difficult To Classify

What can be done if the subject's behavior is not precisely classifiable so that there is an opportunity for subjective observer bias? One wants to determine the pike's preference after it has detected both prey types. We did this kind of experiment with sticklebacks that had the choice- between parasitized and unparasitized copepods that were confined in Plexiglas cells (Wedekind and Milinski, 1996). We wanted to count the fish's bites (snout contacts with the Plexiglas) toward each prey type for one minute. The problem was that the fish could bite at a high rate so that two bites could be regarded as a single bite on the video record taken from the front wall (Fig. 3). Because the prey types could be easily recognized by their behavior, a naive observer would not have solved the problem. We used a second video camera aimed exactly along the front wall of the cells that contained

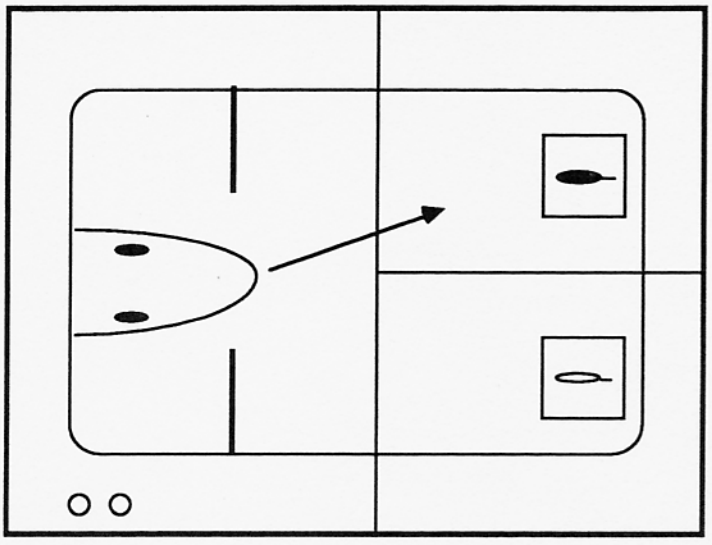

FIG. 2. Video screen showing the experimental setup from above. A pike has entered the experimental side of its tank by a door; when the tip of the pike's snout passes the vertical decision line that is drawn on the screen, it is regarded as having chosen either the conspicuous or the cryptic prey dependent on where the tip of its snout is with respect to the horizontal decision line.

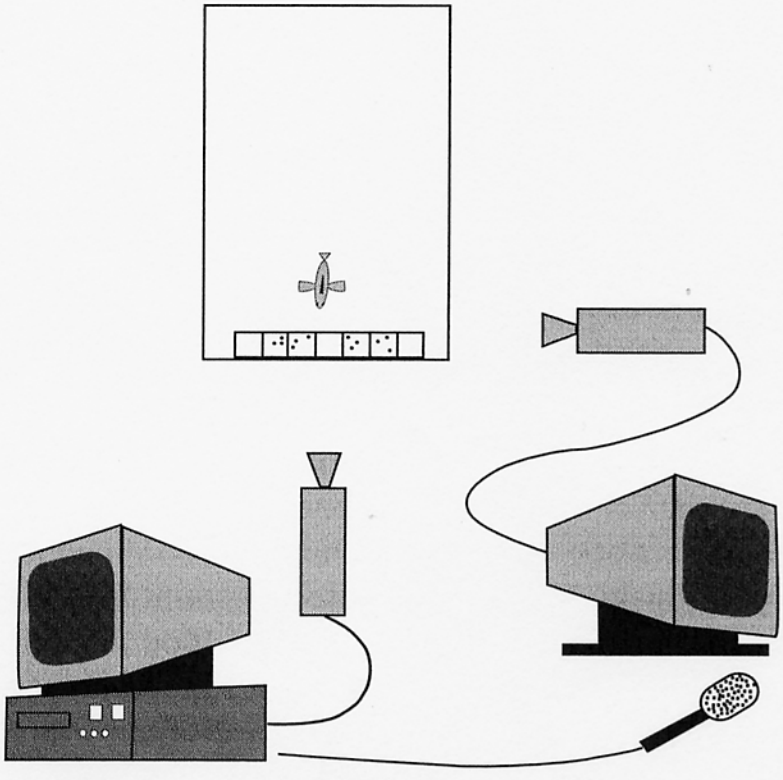

FIG. 3. Experimental setup with which behaviors of the fish can be recorded as bites without the risk of observer bias; see text for further explanations. Drawing by C. Wedekind.

the copepods so that only the bites toward the Plexiglas could be seen and not the type of prey that was attacked. From this view I decided when behaviors were bites and spoke these events into the audio channel of the video record of the fish's choice of prey type (which I could not see). From the latter video and audio record we could determine how often each prey type (as defined by the cell that it was confined in) was attacked by the fish. Sometimes a very elaborate technical design is necessary to remove the opportunity for observer bias. The solution must be tailored specifically to the actual problem so that no cookbook recipe can be offered that fits every situation. It needs creativity to find the solution for specific problems after the researcher has become aware of all opportunities for observer bias in the design.

D. Experimenter Bias

If one does not use a video camera but observes the pike's predation attempts directly, one's presence might affect the behavior of both the pike and the prey. There is a further more subtle opportunity for biasing the results: because one expects the pike to prefer the conspicuous prey one might "hold one's breath" and "freeze" whenever the conspicuous prey is more likely to be attacked. In this way the pike may be less often disturbed when it tries to attack the conspicuous prey, which might therefore inevitably be attacked more frequently. This is similar to the "clever Hans effect": the clever horse appears to be able to count because his master provides him with subtle signs when the correct number is achieved. The use of a video camera with the observer sitting in another room would help to prevent this bias. A blind with a peep hole for the observer would be the second best solution because it still allows for interaction between observer and subject, for example, by vibrations when the observer becomes nervous.

The results may also be biased even when one observes the pike's behavior from another room: each of the two prey types had to be transferred to its Plexiglas tank; each fish had to be caught from its individual tank (with a net), carried to the experimental tank and released into its Plexiglas container. Here, there is ample opportunity for experimenter bias. One may be more careful with the cryptic fish so that they are less frightened than the conspicuous fish in the experiment. The pike may prefer the conspicuous fish because they are more nervous. This problem can be solved by asking another person (who must, of course, be trained to handle live fish carefully) to transfer the fish. However, the naive helper may like the cryptic fish more than the conspicuous fish and may thus treat them differently. Now we are stuck.

It is tempting to suggest that one can always find a simple and elegant

solution. One can toss a coin to decide whether a fish becomes conspicuous or cryptic *after* it has been transferred to its tank! One can position plates behind each Plexiglas tank that either match the conspicuous fish's color (and contrast with the cryptic fish's color) or contrast with the conspicuous fish's color (and match the cryptic fish's color). This kind of technique has been used, for example, by Dawkins (1971). Why not use only one type of fish and affect its conspicuousness only by matching or contrasting backgrounds? Because the results would be ambiguous, since the pike may like one kind of background better.

V. Potential Artifacts When Animals Are Not Accustomed to Experimental Procedures

Whenever an animal is handled (caught, marked, kept under restricted conditions, etc.), this experience will probably affect its future actions. For example, if an animal is frightened by being caught and transferred to a new artificial environment where it is expected to solve a foraging task, it may give priority to avoiding any potential danger instead of to foraging. When it finally starts to forage, it will certainly make a compromise between avoiding potential predation and efficient foraging (e.g., Milinski, 1993). If the experimenter aims to test the predictions of an optimal diet model, he or she might conclude that this species does not select the optimal diet that is predicted. Although this example might appear to be an extreme one, I am convinced that this kind of mistake is made very often. Neither the referees nor the readers of a publication can judge from the Methods section whether the results are flawed by severe handling artifacts. I would like to propose that a detailed description of the methods that were used to habituate the animal to the experimental procedures is required by the "Instructions for Authors" of all behavioral journals. What kind of method should be used to avoid artifacts from handling? Experimenters could find this method if they tried to put themselves in the animal's place to see what happens to it from its point of view. What information is actually available to the animal? What might it mean functionally for the animal?

The only general rule that I can propose is to take the animal through all the experimental steps many times, during the course of several days, except for presenting it with the actual test. For example, being caught and transferred should become a neutral or even a positive event with respect to the task that the animal will be presented with. To illustrate this: I hung the net used to catch my sticklebacks for an experiment in their holding tank every day. Occasionally I offered some food in the net and moved the net with the fish, at first only a bit but later I lifted it out of the tank

for a few seconds (e.g., Milinski, 1985). After several such training sessions the fish swam into the net even when no food was offered.

If the experimental environment is different from its usual environment, the animal should become accustomed to it by being transferred back and forth often enough for it to perceive the experimental environment as familiar with a neutral or positive value. A sentence such as, "after being transferred to the experimental cage, tank, etc., the animal was allowed to become accustomed to it for five minutes before the experiment started," makes me very sceptical about the value of the results. Such a procedure would be justified only if the animal's housing conditions are almost identical to the experimental conditions. The experimenter needs experience with the animal, the necessary skills and intuition to provide the animal with only the test stimulus and not with many strange influences in addition to the experimental ones. In our Example 1 the pike lives in the experimental tank and may be trained to attack fish that are confined in Plexiglas containers; it should receive some other food after each attack, otherwise it would cease to attack unreachable prey. The prey fish will, however, behave abnormally within their containers especially during the pike's approach. They would not only need to be accustomed to this environment but also to be prevented from seeing the pike, for example, by one-way mirrors that allow the pike to see the prey fish but the latter not to see the pike (Magurran, 1990b).

VI. Unsuitable Controls

Ideally, treatment and control differ only in the trait the effect of which is to be tested. In Example 1 different methods can be used to make one prey fish conspicuous and the other one cryptic. If these fish are cryptic anyway it might seem obvious that one should change the appearance of only the fish that are to become conspicuous. One could apply some red dye to the fish's skin. The treatment fish are thus conspicuous and the control fish are cryptic. However, the treatment fish differ not only in conspicuousness from the control fish but also by having been handled and the skin treated by a chemical substance. The control fish would have to be treated in the same way, for example, with green dye, so that any preference must be due to the color only. Now the cryptic fish are an excellent control for the treatment fish. A further problem might be that both groups behave abnormally because their skin is continuously irritated by the chemical substance during the test. The pike may thus not discriminate between them as it would if the fish behaved normally. So this does not work. The solution may be to use dead fish as prey. Their behavior

will not be changed by being handled and treated with dye. They would mimic fish that "freeze" after having detected a predator.

Another solution to our problem with Example 1 would be to change the treatment fish's color through its diet; an example here is male guppies that received supplementary carotenoids with their food so that they became redder than control fish (Kodric-Brown, 1989). How do we treat the control fish? As far as I know, there is nothing comparable to carotenoids that induces a cryptic color in fish. Even if such a substance were available, it would not help much to create a proper control because its effects on the fish's behavior might differ from that of carotenoids. It is not obvious that carotenoids change the behavior of a fish. Can one use just untreated fish as controls? Yes, if another experiment shows that the behavior of fish that were treated with carotenoids does not differ from the behavior of untreated fish. One could determine the antipredator behavior of 10 fish that had been treated with carotenoids and of 10 untreated fish, each fish tested singly after a dummy predator attack. Imagine that we do not find a significant difference (at the level of $\alpha = .05$), but the sample size is much too small for the null hypothesis to be accepted (at the level of $\beta = .20$; see next section). This kind of control experiment is not any help. The sample size would have to be increased enormously, which may not be feasable because it would need more than 10 times the effort needed for the main experiment. This is a real dilemma, which has not yet been appreciated by many researchers.

I propose the following procedure to circumvent the problem of "proving" that a treatment has no undesirable secondary effect. One increases the magnitude of the treatment (or the trait that cannot be excluded) stepwise until there is a significant secondary effect in the treatment group as compared to the control group; one thus tries to demonstrate an effect instead of "proving" that no effect exists, and this needs only small sample sizes. Then one can do a regression analysis to estimate the magnitude of the secondary effect for the treatment in the main experiment. For example, if the basic amount of carotenoids that is used to induce the red color in the conspicuous fish does not obviously induce a change in their antipredator behavior, another group is fed a higher amount and a third group an even higher amount of carotenoids. From this experimentally established relationship between amount of carotenoids consumed and change in antipredator behavior, the behavioral change that is induced by the basic amount of carotenoids can be estimated. With this information it can be discussed whether the potential secondary effect of the basic treatment is of a biologically important size. A problem can arise if even the largest amount of carotenoids that the fish consumed does not induce a recognizable behavioral change. In this case I would be confident that the much

smaller basic amount had no effect, but I cannot really prove it. It would be good if this scenario were to be investigated by statisticians and advice provided for behavioral scientists.

It cannot be guaranteed that a given treatment differs from a given control only in the trait that is under investigation. This ideal difference can be approached only by spending a lot of effort thinking about other potential differences. These should be discussed in the published paper so that the readers can form their own judgment of the suitability of the control.

VII. "Proving" the Null Hypothesis with Small Samples

Sometimes one needs to support the null hypothesis. For example, if treatment and control differ not only in the trait that is under investigation but also in some other trait, we have to "prove" (i.e., reach a threshold of reliability that is set by a convention) that this other trait has no effect (or at most a negligible effect) on the behavior that is measured as a response to the trait under investigation. In our example from the previous section, no significant effect of carotenoids was found on the antipredator behavior of 10 fish in comparison with 10 untreated fish. Can we conclude that the pike prefers fish that were treated with carotenoids because they are redder and not because their antipredator behavior is changed? Did we "prove" the null hypothesis that assumes no effect of carotenoids on antipredator behavior? Is there any convention for accepting the null hypothesis?

"Proving" the Null Hypothesis

Cohen writes in his excellent book on power statistics (1988, p. 16):

> **Research reports in the literature are frequently flawed by conclusions** that state or imply that the null hypothesis is true. For example, following the finding that the difference between two sample means is not statistically significant, instead of properly concluding from this failure to reject the null hypothesis that the data do not warrant the conclusion that the population means differ, the writer concludes, at least implicitly, that there is NO difference. The latter conclusion is always strictly invalid, and is functionally invalid as well unless power is high. The high frequency of this invalid interpretation can be laid squarely at the doorstep of the general neglect of attention to statistical power in the training of behavioral scientists.

Scientists are usually concerned with having a very low significance level α (Type I error), the probability of falsely rejecting the null hypothesis; for example, "the pike significantly ($p < .05$) prefers conspicuous over cryptic prey." However, taking a very small α results in power values being

very small. The complement of power (1 − power) is the β error (Type II error); β represents the error rate of failing to reject a false null hypothesis. So we need a convention about β if we want to accept the null hypothesis as we have with α for rejecting the null hypothesis. Since β is the complement of power, we obtain a smaller β when we increase power. For example, for power = .95, the β error equals .05. How can we increase power? Power is a function of α, effect size, and n (sample size) (Cohen, 1988). Effect size is the difference between treatment and control, which is negligible or 0 if we assume "no difference." If we aim at $\beta < .05$, we have to increase power to .95. For a zero or negligible effect size we have to determine the sample size with which we can achieve a power of .95.

Cohen (1988, p. 17) gives an example. For a correlation analysis, the null hypothesis would be $\rho = 0$, or if the effect size is negligible $\rho = .10$.

> If, for example, one considers a population $r = 0.10$ as negligible (hence, i), and plans a test of the null hypothesis (at $\alpha = 0.05$) for power = 0.95 ($\beta = 0.05$) to detect i, one discovers that the n required is 1308; for power = 0.90 ($\beta = 0.10$), the required $n = 1046$; and for power = 0.80 ($\beta = 0.20$), $n = 783$.

This example shows that, even if we want effectively to "prove" (assuming a small "negligible" effect size instead of zero) the null hypothesis, we need enormous sample sizes for this proof, which are usually not feasible. Cohen (p. 56) proposes as a convention that one sets the power at .80 ($\beta = .20$). This assumes that Type I errors (falsely rejecting the null hypothesis) are of the order of four times as serious as Type II errors (falsely accepting the null hypothesis). Even with this relaxed convention we need sample sizes that are so large that a "proof" of the null hypothesis is probably most often impossible. Everybody who needs to conclude that "no difference between treatment and control existed" should perform a statistical power analysis and include the necessary information in the results section. Any published conclusion of "no effect" or "no difference" without this information should be regarded as unproven. A potential solution to this dilemma is the procedure that I proposed in the previous section.

Of course, much smaller sample sizes are needed to reach the level of $\beta = .20$ if no significant effect can be found and the expected effect size is much larger than "negligible." "All or nothing" responses that appear when the trait has reached a certain threshold are of this kind.

VIII. Conclusions

I hope that researchers will become vigilant about the need to avoid the kind of mistakes that I have discussed in this article. Referees and editors

should ask authors to discuss in detail the methods with which they have avoided these mistakes. As behaviorists we have to work sometimes with limited samples and we cannot have complete control over all conditions. Thus, we have to make compromises. In this case one should be aware of these limitations and be cautious about the conclusions that the study might still allow. I would be glad if readers have become convinced that studying a more detailed review or book on experimental design is worth the investment.

IX. Summary

Seven common mistakes in designing and performing studies of behavior and in interpreting their results are discussed with simple examples: (1) unjustified conclusions are made from observational (i.e., correlational) data; (2) data are not independent ("pseudoreplication"); (3) treatments are confounded by time and sequence effects; (4) no efforts are made to avoid observer bias; (5) potential artifacts arise when animals are not accustomed to experimental procedures; (6) unsuitable controls are used; (7) an attempt is made to "prove" the null hypothesis with small samples.

Acknowledgments

I thank Jay Rosenblatt, Peter Salter, Charles T. Snowdon, and Claus Wedekind for very helpful comments.

References

Andersson, M. (1982). Female choice selects for extreme tail length in a widowbird. *Nature (London)* **299,** 818–820.

Cohen, J. (1988). "Statistical Power Analysis for the Behavioral Sciences," 2nd ed. Lawrence Erlbaum Associates, Hillsdale, New Jersey.

Curio, E. (1983). Time-energy budgets and optimization. *Experientia* **39,** 25–34.

Curio, E., Ernst, U., and Vieth, W. (1978). Cultural transmission of enemy recognition: One function of mobbing. ***Science 202, 899-901.***

Dawkins, M. (1971). Shifts of "attention" in chicks during feeding. *Anim. Behuv.* **19,** 575–582.

Gardner, B. T., and Gardner, R. A. (1969). Teaching sign language to a chimpanzee. ***Science*** **165,** 664–672.

Grafen, A. (1990). Biological signals as handicaps. *J. Theor. Biol.* **144,** 517–546.

Huntingford, F. A. (1976). The relationship between anti-predator behaviour and aggression among conspecifics in the three-spined stickleback, *Gasterosteus aculeatus.* ***Anirn. Behuv.*** **24,** 245–260.

Hurlbert, S. H. (1984). Pseudoreplication and the design of ecological field experiments. *Ecol. Monogr.* **54,** 187–211.

Kodric-Brown, A. (1989). Dietary carotenoids and male mating success in the guppy: An environmental component to female choice. *Behav. Ecol. Sociobiol. 25, 393-401.*

Kroodsma, D. E. (1986). Design of song playback experiments. ***Auk 103, 640-642.***

Kroodsma, D. E. (1989). Suggested experimental designs for song playbacks. ***Anim. Behav.*** **37,** 600–609.

Magurran, A. E. (1990a). The inheritance and development of minnow anti-predator behaviour. ***Anim. Behav.*** **39,** 834–842.

Magurran, A. E. (1990b). The adaptive significance of schooling in an anti-predator defence in fish. ***Ann. Zool. Fennici 27, 51-66.***

Magurran, A. E. (1993). Individual differences and alternative behaviours. *In* "Behaviour of Teleost Fishes" (T. J. Pitcher, ed), pp. 441-477, 2nd ed. Chapman and Hall, London.

Martin, P., and Bateson, P. (1986). "Measuring Behaviour," 1st ed. Cambridge University Press, Cambridge.

Martin, P., and Bateson, P. (1993). "Measuring Behaviour," 2nd ed. Cambridge University Press, Cambridge.

Milinski, M. (1985). Risk of predation of parasitized sticklebacks (*Gasterosteus aculeatus* L.) under competition for food. ***Behaviour 93, 203-216.***

Milinski, M. (1987). TIT FOR TAT in sticklebacks and the evolution of cooperation. *Nature* ***(London)*** **325,** 433–435.

Milinski, M. (1993). Predation risk and feeding behaviour. In "Behaviour of Teleost Fishes" (T. J. Pitcher, ed), pp. 285–305, 2nd ed. Chapman and Hall, London.

Milinski, M., and Bakker, T. C. M. (1990). Female sticklebacks use male coloration in mate choice and hence avoid parasitized males. *Nature* ***(London) 344, 330-333.***

Moller, A. P. (1988). Female choice selects for male sexual tail ornaments in the monogamous swallow. *Nature* ***(London)*** **332,** 640–642.

Moller, A. P. (1989). Viability costs of male tail ornaments in a swallow. *Nature* ***(London) 339, 132-135.***

Moller, A. P. (1990a). Male tail length and female mate choice in the monogamous swallow ***Hirundo rustica. Anim. Behav.*** **39,** 458–465.

Moller, A. P. (1990b). Effects of a haematophagous mite on the barn swallow ***(Hirundo rustica)***: A test of the Hamilton and Zuk hypothesis. ***Evolution*** **44,** 771–784.

Pettifor, R. A. (1990). The effect of avian mobbing on a potential predator, the European kestrel, *Falco tinnunculus.* ***Anim. Behav. 39, 821-827.***

Wedekind, C., and Milinski, M. (1996). Do three-spined sticklebacks avoid to consume copepods, the first intermediate host of ***Schistocephalus soZidus?-An*** experimental analysis of behavioural resistance. *Parasitology.* ***112, 371-383.***

Wedekind, C., Seebeck, T., Bettens, F., and Paepke, A. J. (1995). MHC-dependent mate preferences in humans. *Proc.* ***R.*** *Soc.* ***London B 260, 245-249.***

Wilson, D. S., Clark, A. B., Coleman, K., and Dearstyne, T. (1994). Shyness and boldness in humans and other animals. *Trends Ecol. Evol.* **9,** 442–446.