Seeing What We Want to See: Confirmation Bias in Animal Behavior Research

David M. Marsh & Teresa J. Hanlon

Department of Biology, Washington and Lee University, Lexington, VA, USA

Correspondence
David M. Marsh, Department of Biology, Washington and Lee University, 204 West Washington St., Lexington, VA 24450, USA.
E-mail: marshd@wlu.edu

Received: April 11, 2007
Initial acceptance: June 14, 2007
Final acceptance: July 10, 2007
(S. A. Foster)

doi: 10.1111/j.1439-0310.2007.01406.x

Abstract

Confirmation bias is the tendency of observers to see what they expect to see while conducting scientific research. Although confirmation bias has been well-studied by psychologists in the context of qualitative judgments, it has been much less studied with respect to the kinds of quantitative observations made by behavioral biologists. We carried out two experiments that used multiple observers of the aggression and foraging behaviors of red-backed salamanders (*Plethodon cinereus*) to determine whether behavioral observations were influenced by the a priori expectations of observers. In both experiments, one group of observers was given a specific set of expectations with respect to sex differences in salamander behavior, while a second group was given the opposite set of expectations. In one experiment, observers collected data on variable sets of live salamanders, while in the other experiment, observers collected data from identical videotaped trials. Across experiments and observed behaviors, the expectations of observers did appear to bias observations, but only to a small or moderate degree. Confirmation bias never accounted for more than 13% of the observed variation in behavior, and was generally equivalent to <20% of the mean value of each variable. The estimated magnitude of confirmation bias was quite similar for men and women, suggesting no relationship between observer gender and susceptibility to confirmation bias. We believe that these results are largely optimistic with respect to confirmation bias in behavioral ecology, in that they suggest the bias may often be small relative to individual variation in behavior, even for relatively inexperienced observers.

Introduction

The science of ethology has long contended with biases that may render behavioral observations less than completely objective (Altmann 1974; Jordan & Burghardt 1986; Caine 1990; Kelly 2006). One important type of bias encountered in behavioral research is confirmation bias, also sometimes referred to as confirmatory bias or expectation bias. Confirmation bias refers to the tendency of observers to see what they expect to see when making observations (Rosenthal 1966; Nickerson 1998). Confirmation bias has been well-documented in the human psychology literature, particularly with respect to the way that people's pre-conceptions can influence their subjective judgment about other people (see Rosenthal 1994; Nickerson 1998 for reviews). Confirmation bias may also underlie some of the placebo effect in medicine, in the sense that both patients and clinicians may have a tendency to report the treatment outcomes that they hope to see (Gracely et al. 1985; Kirsch & Weixel 1988; Pollo et al. 2001). Ideally, most quantitative observations of animal behavior are less subjective than the

qualitative judgments common in human psychology research, and thus, should also be less susceptible to confirmation bias. Indeed, placebo effects in medicine have been suggested to be common only for more subjective measures such as pain and discomfort (Hrobjartsson & Gotzsche 2001, 2004). Nevertheless, the extent to which various kinds of behavioral observations are prone to bias is not always obvious a priori (Marsh & Hanlon 2004). Ultimately, designing experiments in which observers are blind to the treatment assignments of their subjects is really the only way to ensure that confirmation bias cannot influence behavioral data.

Unfortunately, the realities of ethological research may often prevent researchers from conducting blind experiments. Many studies necessarily compare the behavior of different sexes, age classes, morphotypes, or species that may be obviously distinct in appearance. When observers are unavoidably aware of the classification of their subjects, it may be difficult for them to make observations that are completely free of confirmation bias. If the magnitude of confirmation bias in behavioral observations is large, it could interfere with our interpretation of research results and our understanding of animal behavior.

Only one study that we know of has directly addressed the question of confirmation bias in animal behavior research. Several decades ago, Rosenthal & Fode (1963) asked 12 student observers to record the success of rats in running a T-maze over the course of 5 d. Observers were told that one group of rats was 'maze-dull' and the other group was 'maze-bright', though in reality all the rats had been randomly selected for the experiment. Rosenthal and Fode reported that observers did in fact find the supposed 'bright' rats to be more successful than the supposed 'dull' rats in running the maze, particularly later in the experiment. However, the mechanism proposed to explain this confirmation bias – differential reinforcement of bright and dull rats by observers – probably would not apply to most behavioral studies because observers usually do not interact with their subjects in ways that would allow reinforcement to be a factor.

In previous research on the biases inherent in the collection of behavioral data on red-backed salamanders, we examined confirmation bias in an indirect manner (Marsh & Hanlon 2004). We attempted to correlate observers' a priori predictions about sex differences in salamander behavior with the data they actually collected. Across a range of aggressive and foraging behaviors, we found little evidence for confirmation bias. However, several factors limited the strength of our tests. First, observers made predictions about expected differences between male and female salamanders, yet each observer collected data on only one salamander. Thus, observers had no opportunity to directly compare the behavior of the two sexes. Secondly, some specific predictions in our previous trials were quite rare (e.g. very few observers expected that female salamanders would be more aggressive than male salamanders). This reduced our power to detect the effects of differential expectations on the data collected.

In the present study, we used two experiments to directly test the effects of confirmation bias on the observation of aggression and foraging behaviors of red-backed salamanders. In the first experiment, multiple observers collected data on one male salamander and one female salamander in both aggression and foraging trials. One group of observers was told that male salamanders were thought to be more aggressive and more active foragers and those female salamanders were thought to be more efficient foragers. The other group of observers was given the opposite set of predictions with respect to sex differences, and neither group was aware that the predictions varied among observers. This experiment thereby evaluated confirmation bias in the context of trials with multiple animals that vary in behavior. In the second experiment, we used videotaped aggression and foraging trials to allow multiple observers to record the behaviors of the same salamanders at the same time. Once again, two groups of observers were given opposite predictions about the expected differences between male and female salamanders. This second experiment, in contrast to the first, allowed us to quantify the magnitude of confirmation bias in situations where there was no true variation in the behavior of the salamanders observed. For each experiment, we estimated the magnitude of confirmation bias from the difference between the observational data collected by the two groups. We also used the interaction between each observer's gender and their a priori prediction to ask whether genders differed in their susceptibility to confirmation bias.

Methods

Study System

Red-backed salamanders are a terrestrial, lungless, plethodontid salamander found in woodlands throughout the eastern USA and Canada. Their social behavior has been well-studied in both the laboratory

Ethology 113 (2007) 1089–1098

and the field (see Jaeger & Forester 1993 for review). Previous studies have determined that both male and female red-backed salamanders defend territories during at least part of the year (Mathis 1990). Aggressive behavior is also displayed by both sexes, and territorial residents may bite intruders or engage in aggressive displays (Jaeger 1984). Foraging behavior in red-backed salamanders has also been detailed in previous studies (Jaeger et al. 1981, 1982). Red-backed salamanders appear to forage primarily by sight, though they can also use smell to locate non-motile prey (David & Jaeger 1981).

We collected sexually mature red-backed salamanders (snout-vent length 3.3–4.3 cm) from the Jefferson National Forest in Giles County, Virginia in Aug. of 2004 for the first set of experiments and in Aug. of 2005 for the second set of experiments. Salamanders from this population have been used in much of the previous behavioral work on red-backed salamanders (e.g. Thomas et al. 1989; Mathis 1990). We sexed salamanders by holding them up to a fiber-optic light and using the presence of pigmented testes or eggs as diagnostic for males and females (Gillette & Peterson 2001). We then placed salamanders in a 15 cm diameter Petri dish lined with filter paper and moistened with declorinated tap water. Experiments described below were carried out in these same Petri dishes. We changed filter paper once per week, at which time salamanders were fed 15–25 wingless *Drosophila melanogaster*. Feeding was stopped 2 wk prior to the start of the experiment to ensure that salamanders would be motivated to forage.

Experimental observers consisted primarily of first and second year university students at Washington and Lee University (Lexington, VA, USA) between the ages of 18 and 21 yr. Of the 186 total observers, 73 were male and 113 were female. We refer to male and female observers as 'men' and 'women', and hereafter restrict the use of the terms 'male' and 'female' to refer to the sex of the salamanders studied. For clarity, we use the term 'gender' for men and women only, and 'sex' for salamanders only.

Observers collected behavioral data on salamanders within the Petri dishes during the day, and video clips were also recorded under ambient light. Red-backed salamanders in nature are primarily active at night and more realistic experiments are carried out in the dark and with less disturbance to focal animals. However, our primary interest was not in the behavior of red-backed salamanders, which has been the subject of numerous, detailed behavioral studies (see Jaeger & Forester 1993; Petranka 1998 for reviews). Rather, we were interested in the susceptibility of the observers to confirmation bias as they recorded their observations. Thus, we assume only that red-backed salamanders exhibit classifiable behaviors under the conditions of the experiment, not that these behaviors are representative of red-backed salamanders in nature or in other laboratory experiments.

Live Interactions

The observations of live salamanders took place from Nov. 15–18, 2004. Observers were divided among five groups that ranged in size from 13 to 17 and there were a total of 74 observers. When observers arrived for the experiment, we asked them to read materials that outlined the experimental procedures. These materials included the following statement about the behavioral differences between male and female salamanders:

> Why would males and females differ? Females carry eggs, and as such, may have different requirements, preferences, and limitations than males. Males of course do not have eggs, but are concerned with getting enough food to last them through the winter, when they'll be largely inactive. Red-backed salamanders are known to defend territories (probably they defend individual rocks or logs), though they rarely fight.

One version went on to state: 'theory suggests that (1) females will be more aggressive than males, (2) males will be more efficient foragers, (3) females will be more active foragers. Our data will allow us to verify these predictions.' In contrast, a second version stated: 'theory suggests that (1) males will be more aggressive than females, (2) females will be more efficient foragers, (3) males will be more active foragers. Our data will allow us to verify these predictions.' Observers were randomly assigned one of these two versions subject to the constraint that equal numbers of each version were used with each group. The statement that 'our data will allow us to verify these predictions' might be construed as particularly likely to elicit confirmation bias. However, it was our goal to evaluate confirmation bias in something approaching a worst-case scenario. For a similar reason, we elected not to use a control treatment in which no biasing statement was given. Previous trials without such a biasing statement found no meaningful behavioral differences between male and female salamanders under experimental conditions (Marsh & Hanlon 2004), so we sought to allocate maximal sample sizes to the variably biased treatments.

At the beginning of each set of trials, we described the relevant aggressive and foraging behaviors to all observers as a group. Prior to initiating observations, observers were asked to note their predictions as to which sex they expected to be more aggressive, to be more efficient foragers, and to be more active foragers. These responses were used to verify that the observers had in fact read and understood the predictions that they had been given. Data from observers whose noted predictions did not match the predictions they had read were eliminated prior to analysis – this occurred in seven cases across all four trials and no more than three times in any one trial.

In the first set of trials, observers attempted to determine whether male or female salamanders would be more aggressive toward an intruder. To measure aggression, observers were asked to record the frequency of four behaviors that may be associated with aggression in plethodontid salamanders. These behaviors likely vary in subjectivity, though all have been used in published works on aggressive behavior in plethodontids (Jaeger 1984; Nishikawa 1987; Lancaster & Jaeger 1995; Rissler et al. 2000; Marsh & Hanlon 2004). The behaviors recorded were:

1. Resident initiates a touch of intruder ('touches').
2. Resident moves toward intruder ('moves').
3. Resident turns to look toward the intruder ('looks').
4. Resident in all-trunk-raised posture ('ATR') , in which its entire trunk is lifted off the substrate.

For these trials, each observer was given four salamanders in separate Petri dishes. Two dishes were labeled 'R' for resident and two were labeled 'I' for intruder. Each observer had one female resident and one male resident, and these were labeled 'F' and 'M'. The sex of the intruder salamander was not specified. Salamanders were randomly assigned to observers and the order in which male and female residents were run was alternated between observers. Observers were asked to record the sex of each resident salamander before each trial to ensure that they were indeed aware of the sex of the salamander being observed. Aggression trials began when observers picked up the first intruder from its dish and placed it in the dish of the resident. Observers then watched the resident salamander for 20 min divided into 40 intervals of 30 s each. Within each 30 s interval, observers recorded whether each of the aggressive behaviors occurred. Observers then ran an identical trial with the resident salamander of the other sex and with the second intruder. Throughout these trials observers handled only the intruders and not the resident animals on which data were collected.

In the second set of trials, observers collected data on foraging behavior. New salamanders were used for the foraging trials and these salamanders were assigned randomly to observers. Again, each observer collected data on one male salamander and one female salamander in a random order. At the beginning of the trial, observers introduced 10 wingless fruit flies (*D. melanogaster*) into the Petri dish containing the focal salamander. Observers were asked to record three foraging behaviors:

1. Time to eat seven out of 10 introduced flies ('time').
2. The number of steps taken with the front forelimbs during the trial ('steps'). We defined a step as any time a forelimb was lifted and then lowered back to the surface of the Petri dish.
3. Lunges made by salamanders in the attempt to capture fruit flies ('lunges'). Observers also recorded whether or not each lunge was successful in capturing a fruit fly.

For the variables 'lunges' and 'steps', observers recorded the total number of behaviors observed until seven of the 10 flies were consumed. If seven out of 10 flies were not consumed within 10 min, the trial was terminated at this point. If salamanders consumed three or fewer flies they were considered non-responsive (n = 10) and eliminated from the dataset. We used the number of steps per minute as an index of foraging activity. We used the proportion of lunges that resulted in the capture of a fly ('accuracy') as an index of foraging efficiency for each salamander.

We used a total of 176 red-backed salamanders in the live interactions, and no salamander was used more than once in each type of trial. Most of the salamanders used in the aggression trials were also used in the foraging trials. For consistency, these salamanders were all used in the aggression trial first, then given at least 48 h before their subsequent use in the foraging trials. Observers were not permitted to talk to one another during or between the trials to avoid comparisons of observations or predictions. Informal questioning of observers after the trials found no evidence that any observers were aware that different observers had been given different predictions for the trials.

Videotaped Interactions

From Oct. 24–27, 2005, we carried out a follow-up experiment using videotapes of red-backed

Ethology **113** (2007) 1089–1098

salamander behavior. Videos were made using a Sony DCR-DVD 403 digital video camera (Sony Electronics, Inc., New York, NY, USA) positioned approx. 0.3 m above each focal salamander. Salamanders were recorded in 20 min aggression trials and 10 min foraging trials as in the first experiment. We selected four taped aggression trials and four taped foraging trials for use in this experiment. These eight trials were selected because salamanders were active and clearly visible for the full time period.

Observers were divided into four groups that ranged in size from 14 to 26, and there were a total of 77 observers. Each observer within a group watched the videotaped trials on his/her own computer monitor, and each group collected data from a different set of aggression and foraging videos. As before, all observers collected data on one male salamander and on one female salamander. In general, the procedure used with the videos was identical to the procedure used for the live interactions. There were, however, a few changes. First, in an effort to obtain more accurate observations, we reduced the number of behaviors recorded. In the aggression trials, observers were not asked to record the variable 'looks toward' as they did in the live interactions. In the foraging trials, observers were not asked to record the number of 'steps', and only foraging efficiency was estimated. Secondly, to avoid confusion with video files, the order in which male and female salamanders were observed was kept constant within a group, though it was alternated from one group to the next. Finally, we note that the aggression variable 'ATR' could not be reliably scored from the videos, as it is normally observed by looking beneath the animal to confirm that its trunk is off the substrate. Nevertheless, we asked observers to attempt to record ATR to test the effects of confirmation bias on a variable that was known to be somewhat subjective.

Data Analysis

We analyzed the data separately for the live interactions and the videotaped interactions, and within each of these, we analyzed the data separately for aggression and foraging trials. For each trial, the experimental unit was the recorded difference between the behavior of the male salamander and the behavior of the female salamander for each observer. In the aggression trials, we divided the number of intervals in which each behavior occurred by the total number of intervals. We then calculated the difference between the observed frequency of the behavior in the male salamander and the frequency in the female salamander. This was performed for 'looks', 'moves', 'touches', and 'ATR' in the live interactions and 'moves', 'touches', and 'ATR' in the videotaped trials. In the foraging trials, we divided the total number of steps by the duration of the foraging bout before calculating each male–female difference. For foraging accuracy, male–female difference was calculated without any adjustment for time.

For each male–female difference, we used a general linear model to test the main effect of the a priori prediction that was given to each observer. For example, we tested whether observers who were told that males were more aggressive recorded a higher male–female difference than observers who were told that females were more aggressive. Each model also included a term for the interaction between the a priori prediction and the gender of the observer. A significant interaction would indicate that one gender was more or less prone to confirmation bias than the other. Finally, for the videotaped trials, we included a block effect for day, as observers watched different sets of videos on different days. For the live trials, we did not include a day effect because salamanders were randomized across days and preliminary analyses confirmed that the day effect did not approach statistical significance in any models ($p > 0.25$ in all cases). Thus, each linear model took the form: sex difference = prediction = prediction × gender + day + error, with day included only in the analyses of the videotaped trials. After eliminating samples in which stated expectations were inconsistent with the treatment group, in which necessary data were missing, or in which salamanders were considered non-responsive, final sample sizes (i.e. total sets of paired observations) were $n = 66$ (error $df = 63$) for the live aggression trials, $n = 53$ ($df = 50$) for the live foraging trials, $n = 71$ ($df = 65$) for the videotaped aggression trials, and $n = 71$ ($df = 65$) for the videotaped foraging trials. For validation purposes, we also ran all analyses on the entire dataset (i.e. with no data eliminated) and our results were qualitatively unchanged.

Because the difference between male and female behavioral frequencies does not necessarily follow a known statistical distribution, we tested the significance of each term by bootstrapping, rather than with conventional statistical tests that require assumptions about the underlying distribution of the response variable. We fit the general linear model (GLM) coefficients for each term in the model using maximum likelihood and used these coefficients, rather than F-statistics, for statistical inference.

We bootstrapped confidence intervals (CI) on GLM coefficients by re-sampling independent variables with replacement. When the 95% CI for a coefficient did not overlap zero, we concluded that each term was statistically significant at the 0.05 level. We also calculated CI in which we adjusted alpha levels for the number of variables measured in each experiment (i.e. $\alpha = 0.0125$ for the live aggression trials, $\alpha = 0.0167$ for the taped aggression trials, and $\alpha = 0.025$ for the of live foraging trials). We suggest that the 95% CI be taken as a more liberal criteria and the adjusted CI used as a more conservative benchmark. In any case, our primary goal was to estimate the magnitude (i.e. effect size) of confirmation bias in each trial, not to test the largely trivial null hypothesis that observers are subject to no confirmation bias whatsoever.

We estimated effect size in three ways. First, we calculated effect size as the mean difference between the treatments as a percentage of each variable. For example, an effect size of 10% would mean that the difference between the groups that were given different predictions (i.e. the magnitude of the confirmation bias) was equivalent to 10% of the mean value of each variable. Secondly, we estimated effect size as the standardized mean difference between treatments, which is calculated as the difference between treatment means divided by their pooled standard deviation (SD). Thirdly, we estimated effect size as the proportion of the sum of squares in each GLM that was explained by the main effect for prediction bias. For the videotaped trials, we first removed day-to-day variation (i.e. variation among the videos themselves) so that the live interactions and the videotaped trials could be directly compared with respect to effect size. We fit models, performed bootstrapping, and calculated effect sizes using MATLAB 7.3 (Mathworks, Inc., Natick, MA, USA).

Results

Confirmation Bias in Live Interactions

For the aggression variables, confirmation bias was never statistically significant, though in three of four cases the direction of the effect was consistent with small-to-moderate confirmation bias (Table 1; Fig 1a). Estimated effect sizes for confirmation bias ranged from −3% for moves to 32% for looks, and confirmation bias explained at most 2.3% of the variation in each response variable (Table 2). For the foraging variables, steps per minute (i.e. foraging activity) was not subject to significant confirmation bias (Table 1; Fig 1b), though difference between treatments was equal to approx. 65% of the mean steps per minute (Table 2). For foraging accuracy (i.e. efficiency), there was a statistically significant confirmation bias at $\alpha = 0.05$ equivalent to approx. 30% of the mean accuracy (Fig 1b). Using an alpha adjusted for multiple response variables ($\alpha = 0.025$), CI just overlapped zero ($\beta = -0.002$ to 0.48) indicating a non-significant result. Observers' predictions explained approx. 1% of the variation in steps per minute and 13% of the variation in foraging accuracy (Table 2).

Table 1: Parameter values and 95% CI for the main effect of confirmation bias

Variable	Trial type	Parameter value	95% CI
Touches	Live	0.009	−0.088 to 0.150
Moves	Live	−0.022	−0.270 to 0.214
Looks	Live	0.083	−0.030 to 0.231
ATR	Live	0.107	−0.078 to 0.379
Accuracy	Live	0.247*	0.021 to 0.443*
Steps/min	Live	0.406	−3.368 to 4.072
Touches	Videotaped	0.014	−0.049 to 0.067
Moves	Videotaped	0.015	−0.057 to 0.098
ATR	Videotaped	0.141**	0.023 to 0.293**
Accuracy	Videotaped	0.004	−0.112 to 0.049

CI, confidence interval; ATR, all-trunk-raised.
*Indicates statistical significance at the 0.05 level.
**Indicates statistical significance after adjustment for multiple response variables.

Confirmation Bias in Videotaped Trials

For the aggression variables, male–female differences in moves and touches were not significantly affected by confirmation bias (Table 1; Fig 1c). Effect sizes were small (2% for touches and 15% for moves) and these explained <0.5% of the variation in each response variable (Table 2). ATR, which was often not clearly visible during videotaped trials, was significantly affected by confirmation bias, with bias equivalent to 17% of the mean ATR. This result remained significant after adjusting for multiple response variables ($\beta = 0.02$–0.38, 98.3% CI). For foraging accuracy, there was no significant confirmation bias (Table 1; Fig 1d). Confirmation bias represented only 0.4% of the mean accuracy and this explained only 1% of the variation in this variable (Table 2).

Observer Gender

Across all variables, no gender by prediction interactions were statistically significant at the 0.05 level,

Ethology 113 (2007) 1089–1098

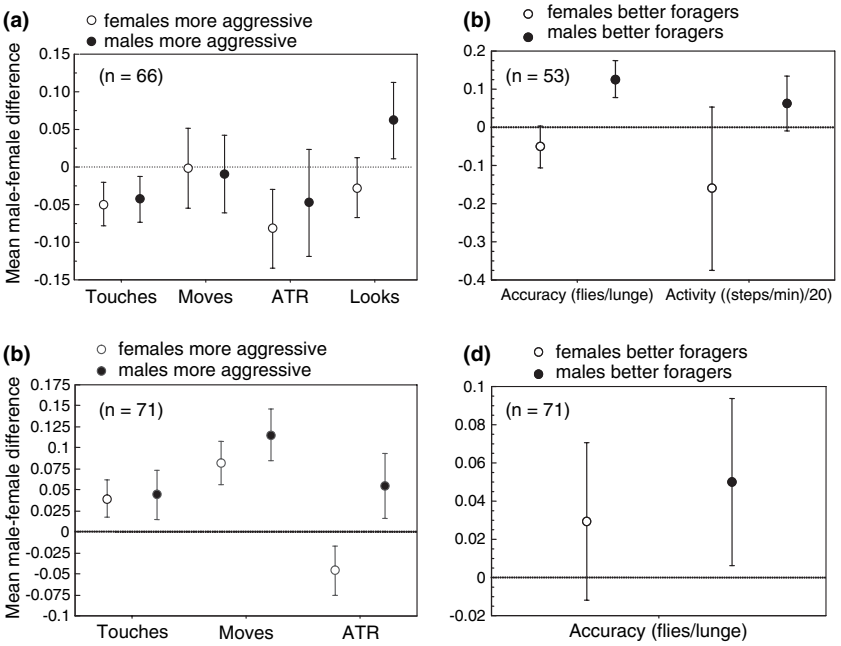

Fig. 1: Confirmation bias in: (a) Live aggression trials. (b) Live foraging trials. (c) Videotaped aggression trials. (d) Videotaped foraging trials. In all cases, male–female differences are predicted to be greater for trials in which observers were told that male salamanders were more aggressive or better foragers (dark symbols) than for trials in which observers were told that females were more aggressive or better foragers (open symbols). In nine of 10 cases, the observed differences were in the direction consistent with confirmation bias, but only foraging accuracy in the live interactions (panel b) and all-trunk-raised (ATR) in the videotaped trials (panel c) were significantly biased.

Table 2: Confirmation bias in the assessment of red-backed salamander aggression and foraging behavior

Variable	Mean difference (%)	Standardized mean difference	Variation explained (%)
Live aggression trials			
Touches	5	0.035	0.3
Moves	−3	−0.026	0.02
Looks	32	0.29	2.3
All-trunk-raised	16	0.10	0.2
Live foraging trials			
Steps/min	65	0.28	1.2
Accuracy	30*	0.61*	13.1*
Videotaped aggression trials			
Touches	2	0.05	0.3
Moves	15	0.33	0.2
All-trunk-raised	17**	0.54**	6.8**
Videotaped foraging trials			
Accuracy	0.4	0.02	1.0

Effect sizes were measured as: (1) the difference between treatment means as a percentage of the mean value of each variable ('mean difference'), (2) the difference between treatment means divided by the pooled SD ('standardized mean difference'), (3) the proportion of the variation in each male–female difference explained by the main effect for confirmation bias ('variation explained').
*Indicates statistical significance at the 0.05 level.
**Indicates statistical significance after adjustment for multiple response variables.

Table 3: Parameter values and 95% CI for interactions between confirmation bias and observer gender

Variable	Trial type	Parameter value	95% CI
Touches	Live	−0.004	−0.149 to 0.107
Moves	Live	0.021	−0.234 to 0.281
Looks	Live	0.001	−0.142 to 0.139
All-trunk-raised	Live	−0.111	−0.365 to 0.076
Accuracy	Live	−0.176	−0.402 to 0.059
Steps/min	Live	−7.36	−39.434 to 1.807
Touches	Videotaped	0.014	−0.039 to 0.074
Moves	Videotaped	−0.044	−0.127 to 0.040
All-trunk-raised	Videotaped	0.058	−0.059 to 0.175
Accuracy	Videotaped	−0.062	−0.160 to 0.020

suggesting that neither gender tended to be more susceptible to confirmation bias (Table 3). Two gender by prediction interactions did approach significance, having model coefficients that just overlapped zero (Table 3). For one of these, men were more affected by confirmation bias whereas for the other, women were more affected.

Discussion

Our results suggest that some confirmation bias was indeed apparent in behavioral observations of red-backed salamanders. For nine of the 10 behaviors recorded, the difference between the prediction treatments was in the direction expected with confirmation bias. In two cases, these differences were statistically significant at the 0.05 level, and in one case this difference was significant after adjusting for multiple response variables. However, the magnitude

of this confirmation bias appeared to generally be small relative to individual variation in behavior. The effect size for confirmation bias ranged from −3% to 65% of the mean value of each variable, with most effect sizes <20%. In terms of variation explained, confirmation bias accounted for <14% of the overall variation in all cases and <3% of the variation in all but the two significant results. Furthermore, the only case of confirmation bias that remained significant after adjustment for multiple response variables was ATR in the videotaped trials, which we knew to be strongly subjective before the trials were run.

Because of the high degree of inter-individual variation in behavior, trials with live salamanders likely had only moderate power to detect significant effects of confirmation bias. Using a post hoc power analysis, we estimate that the magnitude of confirmation bias needed for 80% power would range from 0.06 SD (for looks) to 0.41 SD (for touches). However, in a sense, this low-to-moderate power is exactly the point: because individual variation in behavior was generally large, confirmation bias would have to be quite strong to substantially influence these kinds of results. Certainly our sample sizes, which ranged from 53 to 71 pairs of animals, are not unusually low for a behavioral study. The videotaped trials controlled for this individual variation in behavior and thus yielded higher statistical power. The estimated effect size required for 80% power ranged from 0.03 SD (foraging accuracy) to 0.23 SD (moves).

The obvious question about our research is the extent to which the confirmation biases of university students are representative of the biases of practicing scientists. Undergraduate students are often involved in collecting data for published research, so the distinction between these student observers and practicing scientists is not necessarily as clear as it might seem. In general, though, most behavioral research is conducted by graduate students and PhD-level scientists, and these more experienced researchers might be subject to very different levels of confirmation bias. In general, there are good reasons to view our results as something of a worst-case scenario with respect to confirmation bias. Experienced researchers should be better observers than the naïve observers used in our study, and thus may be subject to even less confirmation bias than we detected. In addition, we used a biasing statement that was quite strong – it suggested that we were examining predictions that were already largely established and that our observations were merely designed to 'verify' these predictions. Finally, we compared trials with opposing biases rather than comparing a biased treatment to a non-biased control. Given this, it could be taken as particularly encouraging that confirmation bias appeared to be generally small in these trials.

There is one important sense, though, in which our results may not represent a worst-case scenario. Our observers had very little stake in the outcome of their observations, as compared with practicing scientists under pressure to publish or get grants. Therefore, it is at least possible that experienced researchers would show more, rather than less confirmation bias, than did our naïve observers. A recent review of confirmation bias in physics suggested that the history of that science is replete with examples of leading researchers being biased toward confirming the predictions of prevailing theories, even when those theories have actually been incorrect (Jeng 2006). Unfortunately, it would be difficult to conduct a research study such as ours with large numbers of PhD scientists who are naïve to the purpose of the study. It may be more feasible to use meta-analysis to evaluate confirmation bias, perhaps by comparing studies where observation were made blind to treatment with similar studies where blinding was not part of the experimental design.

A further question about our research is the extent to which the behaviors of plethodontid salamanders might be subject to more or less confirmation bias than the behaviors of other animals. Some have suggested that many observation biases may be particularly an issue when animals are behaviorally similar to humans, as in the field of primatology (Hrdy 1981; Zuk 2002). Because salamander behaviors are fairly distinct from human behaviors, it may be easier for observers to make unbiased observations on salamanders. However, until we have comparable studies with other taxa, this remains speculative.

Finally, to the extent that we did find some evidence of confirmation bias, the question arises as to the mechanism by which the data may have been biased. Rosenthal & Fode (1963) suggested that the mechanism for confirmation bias in their study of maze running by rats was unconscious reinforcement by observers that differed in magnitude between rats that were falsely labeled 'bright' or 'dull'. In our study, there were no repeated interactions between observers and subjects that could have produced reinforcement. This leaves two main possibilities for the source of the bias. One possibility is that observers could have actually added or

Ethology **113** (2007) 1089–1098

subtracted observations from their original totals in an attempt to produce the predicted results. The other possibility is that observers could have used criteria for scoring behaviors that may have been biased depending on their a priori expectations. We believe that observers had no incentive to knowingly alter their data, and we did not note any erasures or changes in the raw data. We feel the second possibility, that observers scored behaviors differently depending on their expectations, is much more likely. This interpretation is consistent with the fact that the most subjective behavior recorded (ATR in the videotaped trials) was one of the only two statistically significant biases detected.

With respect to observer gender, we found very similar levels of confirmation bias for men and women. Overall, women showed slightly less confirmation bias for six of 10 behaviors, and men showed slightly less bias on four of 10; none of these differences were statistically significant. The subject of gender differences in the way scientific research is conducted is certainly a controversial one, in part because there is little actual data to bear on this issue. In the field of animal behavior, researchers have shown that men and women may tend to choose different study organisms (Holmes & Hitchcock 1997), and may interpret the same data in subtly different ways (Pierrotti et al. 1997). In addition, men and women may have different expectations with respect to sex differences in animal behavior (Marsh & Hanlon 2004). However, given that expectations had only small effects on observed outcomes and that men and women were similarly susceptible to these small biases, we find nothing to suggest that men or women might reach differential conclusions about salamander behavior as a result of confirmation bias. Although our results are somewhat narrow with respect to larger arguments about gender differences in scientific methodology, we find little to support any suggestion of substantial gender differences in the tendency of observers to see what they expect to see.

Acknowledgements

We thank the 2004 and 2005 General Biology laboratory students, the 2004 Animal Behavior students, and the 2005 Behavioral Ecology students for their participation in this research. Mary Jo Kricorian helped set up and carry out the 2005 videotaped trials. Animal care and research methods were covered under IACUC protocol 0204A-DM, and salamanders were collected under Virginia Department of Game and Inland Fisheries permit no. 021066. Use of human subjects was passed by the Institutional Review Board of Washington and Lee University, and all experiments comply with current U.S. Laws.

Literature Cited

Altmann, J. 1974: Observational study of behaviour: sampling methods. Behaviour **49**, 227—267.

Caine, N. G. 1990: Unrecognized anti-predator behaviour can bias observational data. Anim. Behav. **39**, 195—197.

David, R. S. & Jaeger, R. G. 1981: Prey location through chemical cues by a terrestrial salamander. Copeia **1981**, 435—440.

Gillette, J. R. & Peterson, M. G. 2001: The benefits of transparency: candling as a simple method for determining sex in red-backed salamanders (*Plethodon cinereus*). Herp. Rev. **32**, 233—235.

Gracely, R. H., Dubner, R., Deeter, W. R. & Wolskee, P. J. 1985: Clinician's expectations influence placebo analgesia. Lancet **325**, 43.

Holmes, D. J. & Hitchcock, C. L. 1997: A feeling for the organism? An empirical look at gender and research choices of animal behaviorists. In: Feminism and Evolutionary Biology (Gowaty, P. A., ed.). Springer, New York, pp. 184—204.

Hrdy, S. B. 1981: The Woman That Never Evolved. Harvard Univ. Press, Cambridge, Massachusetts.

Hrobjartsson, A. & Gotzsche, P. C. 2001: Is the placebo powerless? An analysis of clinical trials comparing placebo with no treatment. N. Engl. J. Med. **344**, 1594—1602.

Hrobjartsson, A. & Gotzsche, P. C. 2004: Is the placebo powerless? Update of a systematic review with 52 new randomized trials comparing placebo with no treatment. J. Intern. Med. **256**, 91—100.

Jaeger, R. G. 1984: Agonistic behavior of the red-backed salamander. Copeia **1984**, 309—314.

Jaeger, R. G. & Forester, D. C. 1993: Social behavior of plethodontid salamanders. Herpetologica **49**, 163—175.

Jaeger, R. G., Joseph, R. G. & Barnard, D. E. 1981: Foraging tactics of a terrestrial salamander: sustained yield in territories. Anim. Behav. **29**, 1100—1105.

Jaeger, R. G., Barnard, D. E. & Joseph, R. G. 1982: Foraging tactics of a terrestrial salamander: assessing prey density. Am. Nat. **119**, 885—890.

Jeng, M. 2006: A selected history of expectation bias in physics. Am. J. Phys. **74**, 578—583.

Jordan, R. H. & Burghardt, G. M. 1986: Employing an ethogram to detect reactivity of black bears to the presence of humans. Ethology **73**, 89—115.

Kelly, C. D. 2006: Replicating empirical research in behavioral ecology: how and why it should be done but rarely ever is. Q. Rev. Biol. **81**, 221—236.

Kirsch, I. & Weixel, L. J. 1988: Double-blind versus deceptive administration of a placebo. Behav. Neurosci. **102**, 319—323.

Lancaster, D. L. & Jaeger, R. G. 1995: Rules of engagement for adult salamanders in territorial conflicts with heterospecific juveniles. Behav. Ecol. Sociobiol. **37**, 25—29.

Marsh, D. M. & Hanlon, T. J. 2004: Observer gender and observation bias in animal behaviour research: experimental tests with red-backed salamanders. Anim. Behav. **68**, 1425—1433.

Mathis, A. 1990: Territoriality in a terrestrial salamander: the influence of resource quality and body size. Behaviour **112**, 162—174.

Nickerson, R. S. 1998: Confirmation bias: a ubiquitous phenomenon in many guises. Rev. Gen. Psych. **2**, 175—220.

Nishikawa, K. C. 1987: Interspecific aggressive behaviour is salamanders – species-specific interference or misidentification? Anim. Behav. **35**, 263—270.

Petranka, J. W. 1998: Salamanders of the United States and Canada. Smithsonian Institution Press, Washington DC.

Pierrotti, R., Annett, C. A. & Hand, J. L. 1997: Male and female perceptions of pair-bond dynamics: monogamy in western gulls, *Larus occidentalis*. In: Feminism and Evolutionary Biology (Gowaty, P. A., ed.). Springer, New York, pp. 261—275.

Pollo, A., Amanzio, M., Arslanian, A., Casadio, C., Maggi, G. & Benedetti, F. 2001: Response expectancies in placebo analgesia and their clinical relevance. Pain **93**, 77—84.

Rissler, L. J., Barber, A. M. & Wilbur, H. M. 2000: Spatial and behavioural interactions between a native and introduced salamander species. Behav. Ecol. Sociobiol. **48**, 61—68.

Rosenthal, R. 1966: Experimenter Effects in Behavioural Research. Appleton-Century-Crofts, New York.

Rosenthal, R. 1994: Interpersonal expectancy effects: a 30-year perspective. Curr. Dir. Psych. Sci. **3**, 176—179.

Rosenthal, R. & Fode, K. L. 1963: The effect of experimenter bias on the performance of the albino rat. Behav. Sci. **8**, 183—189.

Thomas, J. S., Jaeger, R. G. & Horne, E. A. 1989: Are all females welcome? Agonistic behavior of male red-backed salamanders. Copeia **1989**, 915—920.

Zuk, M. 2002: Sexual Selections. Univ. of California Press, Berkeley, CA.

Ethology **113** (2007) 1089–1098