Comparing Alternative Models to Empirical Data: Cognitive Models of Western Scrub-Jay Foraging Behavior

Barney Luttbeg[1,*] and Tom A. Langen[2,†]

1. National Center for Ecological Analysis and Synthesis, Santa Barbara, California 93109;
2. Departments of Biology and Psychology, Clarkson University, Potsdam, New York 13699

Submitted January 27, 2003; Accepted August 18, 2003; Electronically published February 13, 2004

Online enhancements: appendix tables.

ABSTRACT: Animals often select one item from a set of candidates, as when choosing a foraging site or mate, and are expected to possess accurate and efficient rules for acquiring information and making decisions. Little is known, however, about the decision rules animals use. We compare patterns of information sampling by western scrub-jays (*Aphelocoma californica*) when choosing a nut with three decision rules: best of *n* (BN), flexible threshold (FT), and comparative Bayes (CB). First, we use a null hypothesis testing approach and find that the CB decision rule, in which individuals use past experiences to make nonrandom assessment and choice decisions, produces patterns of behavior that more closely correspond to observed patterns of nut sampling in scrub-jays than the other two rules. This approach does not allow us to quantify how much better CB is at predicting scrub-jay behavior than the other decision rules. Second, we use a model selection approach that uses Akaike Information Criteria to quantify how well alternative models approximate observed data. We find that the CB rule is much more likely to produce the observed patterns of scrub-jay behavior than the other rules. This result provides some of the best empirical evidence of the use of Bayesian information updating by a nonhuman animal.

Keywords: decision rules, diet selection, likelihood statistics, model selection.

Individual animals often must choose one option from an assortment of candidates, as when choosing a mate from an aggregation of advertising suitors, a food item from an assortment of prey, or a territory within a landscape of suitable sites. The decision is typically complicated by uncertainty about the qualities of the candidates and the economics of information gathering to reduce this uncertainty. The processes animals use to make choices under uncertainty are interesting because of what they can tell us about animal cognition and because animal dispersion patterns and mating systems are, in part, a consequence of how animals assess their environment and potential social partners (Abrahams 1986; Abrams 1999; McIntyre and Wiens 1999; Luttbeg, in press).

Several decision rules have been modeled for how individuals can economically and accurately choose a high-quality item from a set of candidates. In this article, we focus on the three rules that have been most frequently invoked as mechanisms of choice under uncertainty: the best of *n* (BN) decision rule, for which a fixed number of candidates (*n*) is sequentially assessed and then the item with the highest apparent quality is selected (Janetos 1980); the flexible threshold (FT) decision rule, for which candidates are sequentially evaluated until one is encountered that surpasses some minimum acceptability criterion, the stringency of which is adjusted downward when the search is proving fruitless (Real 1990); and the comparative Bayes (CB) decision rule, for which searchers have the opportunity to repeatedly assess candidates to improve their estimate of any candidate's quality until the cost of gathering additional information exceeds the expected benefit of improved choice, at which time the candidate that is estimated to be of highest quality is selected (Luttbeg 1996). Each decision rule is intended for a situation in which an individual must choose one member from a pool of candidates, and the searcher can pay a cost to sequentially gather information about members of the candidate pool. Each decision rule also assumes that competition with simultaneously searching rivals is negligible.

Two general approaches have been used to evaluate which decision rule is a better representation of how animals make choices. The first approach has been to compare the performance of decision rule models via simu-

* Present address: Department of Environmental Science and Policy, University of California, Davis, California 95616; e-mail: btluttbeg@ucdavis.edu.

† E-mail: tlangen@clarkson.edu.

Am. Nat. 2004. Vol. 163, pp. 263–276.

lations of individuals choosing options and to infer that the model that provides the highest average payoff is most likely to occur in nature (Janetos 1980; Real 1990; Luttbeg 1996, 2002). However, the performances of decision rules can be very sensitive to assumptions about constraints and functional relationships. For example, Real (1990) demonstrated that, under a general set of assumptions, a threshold decision rule produces a higher average payoff than a BN rule when there are sampling costs. Empiricists used this result to justify discounting the BN rule as a plausible decision rule when there are measurable costs to sampling candidates (e.g., Dale et al. 1992; Gibson 1996; Reid and Stamps 1997). However, Luttbeg (2002) demonstrated that the BN rule outperforms an FT rule despite high assessment costs under certain plausible constraints. More generally, it is probably unwise to discard assessment models simply because their simulated performance is suboptimal; too little is known about mechanisms of animal cognition to be confident that putative "best" rules prevail in nature.

The second approach for evaluating putative decision rules has been to identify distinctive patterns within search sequences that rules produce and to compare these patterns with the patterns found in actual animal search sequences (Wittenberger 1983; Gibson and Langen 1996). However, most sequence patterns are not unique to one decision rule. Moreover, the sequences a particular decision rule produces may depend on the candidate pool size, prior information about candidate pool quality, and assessment accuracy (Valone et al. 1996; Wiegmann et al. 1996; Reid and Stamps 1997; Uy et al. 2001; Luttbeg 2002). For example, one pattern that has been believed to be diagnostic of a particular decision rule is the position of the chosen item in the sequence of sampled items: any item within a sequence is equally likely to be selected if a BN rule is used, but only the last item of the sequence is selected if a fixed threshold rule is used (Bensch and Hasselquist 1992; Fiske and Kalas 1995; Rintamaki et al. 1995; Dale and Slagsvold 1996; Uy et al. 2001). However, this distinction is weakened if individuals use a flexible threshold (Bensch and Hasselquist 1992; Wiegmann et al. 1996) or if individuals using a fixed threshold rule occasionally accept items that they previously rejected because of imperfect information (Luttbeg 2002).

In this article, we apply two different statistical approaches to evaluate which of the three putative decision rules best predicts the nut-sampling patterns of western scrub-jays (*Aphelocoma californica*) when selecting a food item. Our intent is to infer the cognitive processes by which scrub-jays select a nut by comparing empirical sampling sequences to those produced by decision rule models and to demonstrate the advantages of model selection based on likelihoods and information theory. We first use a null hypothesis testing approach to examine whether the decision rules produce the statistically significant patterns that have been observed in scrub-jay nut-sampling sequences, and we discuss some of the limitations of this approach. We then apply a model selection approach based on information theory and examine the evidence the scrub-jay data provide for the alternative models.

Scrub-Jay Nut-Sampling Sequences

Western scrub-jays in California store up to 5,000 acorns in autumn for later retrieval, typically selecting one nut at a time to cache (Curry et al. 2002). Scrub-jays will similarly store whole peanuts presented at feeders—preferentially choosing the heaviest one encountered—by sequentially handling in the beak, one at a time, as many as 25 different peanuts while repeatedly returning to a subset of these to handle again before selecting one peanut to store (Langen and Gibson 1998; Langen 1999). Langen and Gibson (1998) concluded that repeated handling of peanuts provides information on the amount of food contained within, a form of information sampling. Langen (1999) provided evidence that jays dynamically adjust the patterns of sampling depending on candidate pool composition and sampling costs and that information sampling increases the long-term rate of food storage over that of a nonsampling bird.

The peanut-sampling sequence data we analyze in this article are from experiments detailed in Langen (1999). Scrub-jays were presented two, six, or unlimited (1,000–1,200) peanuts on an elevated feeder platform, and a video record was made of the sampling sequences of jays selecting an item. The two- and six-peanut presentations included two different treatment types: a low population variance (all peanuts similar in mass, ratio of the mass of the heaviest peanut and the lightest peanut = 1.2) and high population variance (peanuts include a large range of masses, ratio of the mass of the heaviest peanut and the lightest peanut = 2.1). Thus, these experiments provided sampling sequence data from five different types of candidate pools.

We use the terminology established in Langen (1999) to describe features of the sampling sequences. To describe a sequence, we designate a unique letter label for each sampled peanut, named sequentially in order of handling, with the last letter corresponding to the selected item, for example, ABAAC. In this sequence, the scrub-jay sampled three items (i.e., the number of different peanuts handled), took five samples (i.e., instances of peanut handling), and chose peanut C. A repeat occurs when a jay handles the same peanut consecutively, for example, A<u>A</u>. A switch occurs when the next peanut handled is different from the previous, for example, A<u>B</u>. A return is a switch to a peanut

handled earlier in the sequence, for example, AB<u>A</u>. The sequence ABA is also a special case for which the return is to the most recently sampled before the switch (this is not true for sequence ABC<u>A</u>). We assume that each sample provides information for making a choice, including the last one that results in removal; we cannot know whether a jay has already resolved to select the last-sampled peanut before it is picked up or, alternatively, whether it acquires information during the last pickup that clinches the decision, but we assume the latter.

Langen (1999) identified eight statistically significant characteristic patterns in the sequences of scrub-jays sampling peanuts. (1) Scrub-jays sample more peanuts and (2) make fewer samples per sampled item as the pool of available peanuts increases. (3) Jays sample more peanuts when the population variance in peanut mass is low. (4) When a jay returns to a previously sampled peanut and there is more than one item to which it can potentially return, it returns more often to the peanut handled immediately prior than to other peanuts handled earlier in the sequence. For example, the pattern ABCB is more common than ABCA. (5) Scrub-jays are more likely to return or (6) repeat during their final sample than earlier in the sampling sequence. Thus, ABCB (a return) is more common than ABAC, and ABCC (a repeat) is more common than ABBC. (7) Scrub-jays are more likely to select the peanut that is the last new item to be sampled within a sequence. For example, in the sequence ABCB, C is the last new item. (8) Finally, scrub-jays are more likely to chose the peanut that is sampled the most times than expected by chance. In the sequence ABCB, B is the most frequently sampled item.

Three Alternative Decision Rules

We assume that individuals attempt to maximize the net benefit received from their choice, which for scrub-jays is the mass of a chosen peanut, a quantity that is highly correlated with the amount of food contained within the nut (Langen and Gibson 1998) minus accrued assessment costs. Each time a jay samples an item, it pays an assessment cost. These costs accrue linearly, as would be expected for costs associated with expended energy or time. We also assume that jays can only conduct a maximum of 30 samples, after which a peanut must be selected. We include this constraint because, with finite items and infinite time, application of one of the three decision rules (the FT rule) can result in endless, fruitless searches for a nonexistent item (Luttbeg 2002). We believe that this limit on search length has no effect on our conclusions, because for most parameter values it was rarely reached. We will now summarize the three decision rules and the assumptions of the rules, but a more thorough description of the rules is given in Luttbeg (2002).

Comparative Bayes (CB) Decision Rule

The CB decision rule, as applied to scrub-jay nut choice, posits that a jay during each step in a search will sample the peanut that is expected to provide the most valuable information. However, if the cost of assessment exceeds the expected value of additional information, the jay is predicted to terminate the search and select the peanut with the highest estimated mass at that point (Luttbeg 1996). The CB decision rule assumes that jays have a prior estimate (in the form of a Gaussian distribution) of the mass of each peanut. If a jay has no prior knowledge about the mass of a peanut, its prior estimate of a peanut's mass equals the population distribution of peanut masses. The new information a jay receives when it samples a peanut is combined with its prior estimate of the peanut mass using Bayesian updating to form a posterior estimate. The estimate is shifted toward the new information, and the uncertainty in the estimate is reduced. Key differences between the CB decision rule and the two decision rules described below are that jays treat information gained from sampling as imperfect (e.g., handling a peanut provides only an approximate indication of the true mass of food contained within), use the information provided by sampling to update their estimate of the mass of the peanut, and prospectively choose which peanut to assess next on the basis of the updated estimates.

Best of n *(BN) Decision Rule*

The BN decision rule, as applied to scrub-jay nut choice, posits that a jay will sequentially handle a fixed number (n) of peanuts and then select the peanut from within the pool of sampled peanuts that has the highest apparent mass. Jays set the number of peanuts (n) they sample on the basis of their estimate of the variance in the distribution of peanut masses and the cost of sampling peanuts. We use normal order statistics to find these optimal n's. Following the common formulation of the BN decision rule (Janetos 1980; McKenna 1985; Real 1990), we assume that jays randomly encounter peanuts and treat samples as providing perfect information about a peanut's mass.

Flexible Threshold (FT) Decision Rule

The FT decision rule, as applied to scrub-jay nut choice, posits that a jay sets a threshold for the minimum acceptable mass of a peanut on the basis of its estimate of the mean and variance in the distribution of peanut masses and the cost of sampling peanuts, and it sequentially han-

dles peanuts until one meeting or surpassing the criterion is encountered. As the search for an acceptable peanut continues and the time limit for making a choice approaches, the jay adjusts the threshold downward. We use a dynamic state variable model to find the optimal thresholds (Clark and Mangel 2000). Following the common formulation of the FT decision rule (Real 1990), we assume that jays randomly encounter peanuts and treat samples as providing perfect information about a peanut's mass.

First Approach: Null Hypothesis Testing

Langen (1999) suggested that the statistically significant patterns in scrub-jay peanut-sampling behavior are similar to those produced when a CB decision rule is used (Luttbeg 1996). The null hypothesis testing approach, which is the most common approach in the biological sciences, compares data with a single null hypothesis by calculating the probability of producing data as extreme as or more extreme than observed given that the null hypothesis is true. If the probability of producing the data given the null model is sufficiently small (typically <.05), then the null hypothesis is rejected. This approach is backed by the philosophical stance that through a combination of critical experiments and confrontations between the resulting data and null hypotheses, some hypotheses will be falsified, and thus the remaining hypotheses will have received some provisional confirmatory support from the data (Platt 1964).

This approach is not well suited for model selection based on observed data (Burnham and Anderson 2002). If alternative models are nested (i.e., simpler models match more complex models with parameters eliminated or set to 0), then methods like stepwise regression might be used to investigate how complex the models should be. This approach, however, cannot be used when models, such as ours, are not nested, and the approach tends to overfit the data and does not quantify model selection uncertainty, thus giving the false impression that the "true" model has been found (Burnham and Anderson 2002). An approach that has been used for nonnested models is to examine whether simulations of choice behavior produce the same patterns as the observed data (Mitchell 1975; Brown 1981; Bertorelle et al. 1997), and this is the approach that we take. We do this to highlight some of the limitations of the null hypothesis testing approach: alternative models cannot be directly compared, and the magnitude of support that data provide a model cannot be quantified.

For each of three decision rules, we produce 1,000 replicate runs of an individual choosing a food item. In each run, individuals choose one peanut from a pool of 20; the population distribution of peanut mass has a mean of 1.55 and variance of 0.25. The cost of sampling a peanut is 0.025, and the information received is imperfect with a signal variance of 0.25. In simulations of all three decision rules, individuals receive a signal of a peanut's mass when they sample it. The magnitude of the signal is drawn from a Gaussian distribution with a mean equal to the actual mass of the peanut and a variance equal to the signal variance (Luttbeg 1996). When the signal variance is 0.00, the signal always matches the actual mass of the peanut. As signal variance increases, the signal becomes a less reliable indication of the actual mass of the peanut. To mimic Langen's (1999) manipulations of pool size and the variance in peanut mass, we vary the pool size from the baseline value of 20 peanuts to six and 50 peanuts and vary the population variance in peanut mass from the baseline value of 0.25 to 0.1, 0.5, and 1.0. For each replicate, we form the pool of candidates by randomly drawing the mass of each peanut from the Gaussian distribution of peanut mass. When we manipulate the population variance in peanut mass, either we assume that individuals detect that the variance has changed—causing them to alter their prior estimates if using the CB decision rule, the thresholds if using the FT rule, or the n if using the BN rule—or we assume they do not detect that the variance has changed and leave the decision rules unaltered. The patterns within the simulated sampling sequences are analyzed in the same fashion as described in "Scrub-Jay Nut-Sampling Sequences."

Results of Null Hypothesis Testing

As pool size increased, there was a statistically significant increase in the number of items sampled for each rule (table 1; regression: CB, $r^2 = 0.02$, $P < .05$; BN, $r^2 = 0.63$, $P < .05$; FT, $r^2 = 0.05$, $P < .05$) and a statistically significant decrease in the number of samples per sampled item (regression: CB, $r^2 = 0.01$, $P < .05$; BN, $r^2 = 0.53$, $P < .05$; FT, $r^2 = 0.14$, $P < .05$). When we altered the variance of the peanut masses in a candidate pool and stipulated that the jays detected the change and altered their decision rules accordingly, for each of the three decision rules there was a statistically significant decrease in the number of items sampled as the population variance increased (table 1; regression: CB, $r^2 = 0.15$, $P < .05$; BN, $r^2 = 0.76$, $P < .05$; FT, $r^2 = 0.11$, $P < .05$). This pattern is opposite of the empirical sequence data. When we stipulated that the jays did not detect the change in the variance, as the population variance increased for the CB and FT decision rules, there was a statistically significant increase in the number of items sampled, but for the BN rule, there was no statistically significant change (regression: CB, $r^2 = 0.07$, $P < .05$; BN, $r^2 = 0.00$, $P > .05$; FT, $r^2 = 0.06$, $P < .05$).

When an individual returned to a previously sampled

Table 1: Sequence patterns of each of the three decision rules compared with the characteristic western scrub-jay peanut-sampling sequence patterns

Characteristic sampling sequence patterns	Comparative Bayes	Flexible threshold	Best of *n*
1. Items sampled increases as pool size increases	+	+	+
2. Samples per sampled item decreases as pool size increases	+	+	+
3*a*. Items sampled increases as pool variance decreases, searcher knows pool variance	−	−	−
3*b*. Items sampled increases as pool variance decreases, searcher does not know pool variance	+	+	0
4. Pr(return to immediately prior | return) > random	+	0	0
5. Pr(repeat | final) > Pr(repeat | not final)	+	−	−
6. Pr(return | final) > Pr(return | not final)	+	0	+
7. Pr(last sampled chosen) > random	+	+	0
8. Pr(most sampled chosen) > random	+	+	+

Note: Each sequence pattern tested using the null hypothesis testing approach. A plus sign indicates a statistically significant pattern in agreement with the empirical data, a minus sign indicates a statistically significant pattern in the opposite direction of the empirical data, and a 0 indicates no statistically significant pattern. Scrub-jay data derived from Langen (1999).

item, the frequency of that return being to the item sampled immediately prior was significantly greater than randomly expected for the CB decision rule but was not significantly different from the random expectation for the BN and FT rules (table 1; χ^2 goodness of fit: CB, $\chi^2 = 70.39$, $P < .05$; BN, $\chi^2 = 0.17$, $P > .05$; FT, $\chi^2 = 2.50$, $P > .05$). The probability of a repeat was significantly larger if it was the final sample in a search sequence than earlier within a sequence for the CB rule (χ^2 analysis: CB, $\chi^2 = 62.79$, $P < .05$). For the BN and FT rules, the opposite pattern occurred: the probability of a repeat was significantly smaller if it was the final sample (χ^2 analysis: BN, $\chi^2 = 24.44$, $P < .05$; FT, $\chi^2 = 4.68$, $P < .05$). The probability of a return was significantly larger if it was the final sample in a search sequence than earlier within the sequence for the CB and BN decision rules but not for the FT rule (paired *t*-test: CB, *t* ratio = −5.86, df = 18, $P < .05$; BN, *t* ratio = −10.61, df = 4, $P < .05$; FT, *t* ratio = 1.61, df = 15, $P > .05$). The last new item to be sampled within a sequence was selected significantly more often than randomly expected for the CB and FT decision rules but not for the BN rule (χ^2 goodness of fit: CB, $\chi^2 = 1{,}900.72$, $P < .05$; BN, $\chi^2 = 1.52$, $P > .05$; FT, $\chi^2 = 1{,}201.92$, $P < .05$). The most frequently sampled item was selected significantly more often than randomly expected for all three decision rules (χ^2 goodness of fit: CB, $\chi^2 = 365.04$, $P < .05$; BN, $\chi^2 = 1{,}940.81$, $P < .05$; FT, $\chi^2 = 15.98$, $P < .05$).

In summary, the CB decision rule matched eight of the eight empirical patterns. In comparison, the FT decision rule matched five of the eight empirical patterns and had one case of a statistically significant pattern in the direction opposite of what was observed. The BN decision rule matched four of the eight empirical patterns and also had one case of a statistically significant pattern in the direction opposite of what was observed. Note that we ignore in this tally the mismatches that occurred when it was assumed that jays detected changes in the variance in quality of candidate pool members, for which all three decision rules mismatched, since the alternative plausible formulation positing that jays do not account for changes in variance did produce an empirical match for two of the three decision rules.

These results might be interpreted as evidence that scrub-jays use a decision rule that is more similar to the CB decision rule than to the BN or FT rules. However, this inference suffers from three important weaknesses. First, the patterns of statistically significant results for the empirical and simulation data depend critically on the level of significance (α) at which we choose to declare statistical significance (in these analyses, $P < .05$, as is customary) and on the sample sizes. If we increased the sample sizes of our simulation data, some of the patterns that were not statistically significant may have become so, thus altering our count of how many observed patterns were matched by the rules. Second, our analyses were done with a particular set of parameter values. A sensitivity analysis wherein we systematically varied parameter values and judged the robustness of patterns would provide more confidence in our conclusions (Gladstein et al. 1991). Third, these analyses provide only dichotomous judgments about whether a particular decision rule produces a particular pattern (the simulation pattern is/is not significantly different from random, is/is not in the direction of the empirical pattern). Using this approach, we cannot measure how well each of the decision rules fits the empirical data. For example, cases where each of the three decision rules produces patterns consistent with an ob-

served pattern are uninformative about which rule best matches the data.

Second Approach: Likelihoods of Rules Given Observed Behavioral Patterns

Next, we apply a model selection approach that is based on information theory. The approach requires an empirical data set and a set of a priori alternative models and uses Akaike Information Criteria (AIC) to estimate the Kullback-Leibler distance between the data and each alternative model. This distance represents how well each model approximates the information present within the empirical data set and quantifies the evidence in the empirical data for each of the alternative models (Hilborn and Mangel 1997; Burnham and Anderson 2002). Other criteria exist, and their creation and evaluation are an active area of research (Pitt and Myung 2002; Taper and Lele, in press), but some of them, such as Bayesian Information Criteria and minimum description lengths, do not estimate Kullback-Leibler distances and require large sample sizes to work effectively (Burnham and Anderson 2002). This model comparison approach is situated within the philosophy of simultaneously testing multiple working hypotheses, rather than having a single null hypothesis to compare with a single alternative hypothesis (Chamberlin 1890; Hilborn and Mangel 1997; Anderson et al. 2000). It solves many of the problems associated with the null hypothesis testing approach: alternative hypotheses can be directly compared, and the relative evidence provided by the data for each alternative hypothesis can be quantified (Hilborn and Mangel 1997; Burnham and Anderson 2002).

The first step in this approach is to measure the likelihood of each alternative model—in this case, the three decision rules—producing the data. Typically, one has an underlying probability distribution that can describe the distribution of data a model is expected to produce. However, given the complexity of our alternative decision rules and the behavioral patterns to which they are being compared, there are no obvious probability distributions for describing these expectations. Thus, we use simulations of the decision rules to quantify the probability distributions of patterns formed by each decision rule. This approach is labor intensive but powerful because likelihoods can be calculated for any model that can be simulated.

The simulations are similar to those described above. From the scrub-jay nut-sampling sequence data, we exclude trials with pools of two peanuts, because many of the observed sequence patterns cannot occur when only two items are presented. Thus, we run simulations in which individuals choose from six-item pools with either high or low population variance or from 30-item pools. Simulations with 30-item pools mimic the empirical unlimited peanut experimental trials. It is necessary to limit the unlimited trial type to 30 items to make computation manageable, and 30 seems reasonable since the maximum number of peanuts sampled by scrub-jays in the unlimited trials was 25 (Langen 1999). The mass of each item in a pool is drawn from a Gaussian distribution with a mean of 1.55 g and variances of 0.012 (six items, low variance), 0.14 (six items, high variance), and 0.25 (30 items). These parameters closely match the values in the scrub-jay trials (Langen 1999).

For each decision rule, we systematically vary the cost of sampling from 0.01 to 0.12 at intervals of 0.01, signal variance (σ^2) from 0 to 0.5 at intervals of 0.05, the individual's estimate of the population variance (ρ^2) from 0.025 to 0.475 at intervals of 0.05, and the individual's estimate of the population mean (μ) from 1.30 to 1.80 at intervals of 0.05. For each combination of these parameters and for each decision rule, we run 1,000 replicates of an individual choosing an item. For each replicate, a new set of items is drawn at random from the specified distribution. We record the sequence of items sampled and the item chosen.

The sampling sequences for each of the decision rules depend on the parameter values representing the cost of sampling, the reliability of the information provided by each sample, and the searcher's estimate of the mean and variance of the distribution of pool quality. In some cases, these parameters alter the decision rules. For example, as the cost of assessment increases, the optimal n for the BN rule and the optimal threshold for the FT rule both decline. In other cases, the parameters do not alter the rules but do alter the sampling sequence patterns. For example, the thresholds of the FT rule are unaffected by signal variance, but as signal variance increases, individuals using FT become more likely to end a sampling sequence with a return. Thus, in some cases where a parameter is not in a rule, we still vary the parameter because it affects how well the rule fits the data. For the BN rule, we do not vary the individual's estimate of the population mean because it has no effect on the behavior of individuals using BN. Thus, we fit four parameters for the CB and FT rules and three for the BN rule.

From the sampling sequence patterns produced by the simulations, we extract patterns that as close as possible match the characteristic nut-sampling sequence patterns of the scrub-jays. For the first three characteristic scrub-jay sampling sequence patterns (table 1), we analyze the two statistically independent patterns: the number of samples and the number of items sampled. We compare the empirical distribution of samples and items sampled with distributions formed by the simulations. We calculate the probability that an individual would sample x items given

the treatment t (i.e., six-item pools with either high [6H] or low [6L] population variance or from 30-item [30] pools) as

$$p_{x,t} = \frac{\text{observations of } x \text{ sampled}}{\text{total observations}}. \quad (1)$$

Data from the 30-item treatment (both empirical and simulation) are lumped into categories (y) because of low sample sizes for long sequences (see app. A in the online edition of the *American Naturalist* for a summary of the scrub-jay nut-sampling sequence data).

The order in which the data are reproduced can be ignored since the effect of incorporating the order would be the same for each decision rule, and multiplying the probabilities by a constant has no effect on subsequent analyses. Thus, the probability of a decision rule with a specified set of parameters producing the empirical data is the product of the probabilities (for 6H, six items were never sampled):

$$\Pr(\text{sampled data}|\text{rule and parameters}) = \prod_{x=1}^{6}\prod_{z=1}^{\text{obs}(x)} p_{x,6\text{L}} \prod_{x=1}^{5}\prod_{z=1}^{\text{obs}(x)} p_{x,6\text{H}} \prod_{y=1}^{11}\prod_{z=1}^{\text{obs}(y)} p_{y,30}. \quad (2)$$

The likelihood of the empirical data, given a decision rule and a specified set of parameters, is proportional to the probability of observing the data given a rule and a set of parameters (Hilborn and Mangel 1997):

$$\mathcal{L}(\text{rule and parameters}|\text{sampled data}) = c\Pr(\text{sampled data}|\text{rule and parameters}). \quad (3)$$

We fix $c = 1$, since we are concerned with the relative likelihoods of the decision rules (Hilborn and Mangel 1997). We use the same procedure to compare the number of samples per search sequence with the simulation results from the three decision rules.

For the fourth characteristic sequence pattern, we measure the likelihoods of each decision rule producing the observed number of returns to items sampled immediately prior given a return. We use this metric because the original sequence pattern analysis comparing the probabilities of returns to immediately prior option with other previously sampled options does not produce a distribution of results. Pr(return to immediately prior|return) depends on the number of items sampled to that point in a search sequence. For example, if only two items have been sampled, Pr(return to immediately prior|return) can only equal 1.0. We total the number of returns within sequences and the number of instances that a return was to the immediately prior item, categorize these by the number of items that have been sampled at the point of each return, and calculate $p_{x,t}$, which is the Pr(return to immediately prior|return) for a given number of items sampled (x) and trial type (t), as the number of observed returns to the prior item divided by the total number of returns.

For the six-item trial types, we limit our analyses to instances where three to five items have been sampled, because when only two items have been sampled, returns can be only to the item sampled immediately prior, and returns in which six items had been sampled were rare in the empirical data. For similar reasons, we limit our analyses of the 30-item simulations to instances where three to six items have been sampled. The probability that a rule with a given set of parameters will produce the number of returns to the immediately prior item (u) as observed in the empirical data is

$$\Pr(\text{return data}|\text{rule and parameters}) = \prod_{x=3}^{5}\text{Bin}(u, p_{x,6\text{L}}) \prod_{x=3}^{5}\text{Bin}(u, p_{x,6\text{H}}) \prod_{x=3}^{6}\text{Bin}(u, p_{x,30}). \quad (4)$$

.

For the fifth characteristic sequence pattern, we measure the likelihoods of each decision rule producing the observed number of repeats for final samples and samples earlier within a sequence. We calculate $p_{\text{trial type},l}$ (the probability of a repeat given the trial type and the location in the sampling sequence [l]; either the final [F] sample or not [NF]) as the number of observed repeats divided by the number of opportunities for repeats. The probability that a rule with a given set of parameters will produce the number of repeats (r) observed in the empirical data is

$$\Pr(\text{repeat data}|\text{rule and parameters}) = \text{Bin}(r, p_{6\text{L,F}})\text{Bin}(r, p_{6\text{L,NF}})\text{Bin}(r, p_{6\text{H,F}}) \times \text{Bin}(r, p_{6\text{H,NF}})\text{Bin}(r, p_{30,\text{F}})\text{Bin}(r, p_{30,\text{NF}}). \quad (5)$$

.

For the sixth characteristic sequence pattern, we measure the likelihoods of each decision rule producing the observed number of returns for final samples and samples earlier within a sequence. We calculate $p_{x,\text{trial type},l}$ (the probability of a return given the number of items sampled [x], trial type, and the location in the sampling sequence [l]) as the number of observed returns divided by the number of opportunities for repeats. We include the number of items sampled because the random probability of a return increases as the number sampled increases. The probability that a decision rule with a specified set of parameters will produce the number of returns (n) observed in the empirical data is

$$\Pr(\text{return data}|\text{rule and parameters}) = \prod_{x=2}^{5}\text{Bin}(n, p_{x,6\text{L,F}})\prod_{x=2}^{5}\text{Bin}(n, p_{x,6\text{L,NF}})\prod_{x=2}^{5}\text{Bin}(n, p_{x,6\text{H,F}}) \quad (6)$$
$$\times \prod_{x=2}^{5}\text{Bin}(n, p_{x,6\text{H,NF}})\prod_{x=2}^{8}\text{Bin}(n, p_{x,30,\text{F}})\prod_{x=2}^{8}\text{Bin}(n, p_{x,30,\text{NF}}).$$

For the seventh characteristic pattern, we measure the likelihoods of each decision rule producing the observed number of last newly sampled items being chosen. We calculate $p_{\text{trial type}}$ (the probability of choosing the last sampled item given the trial type) as the observed times the last newly sampled item was chosen divided by the opportunities for it to be chosen. The probability that a decision rule with a specified set of parameters will produce the characteristic pattern is

$$\Pr(\text{choose last data}|\text{rule and parameters}) = \text{Bin}(n, p_{6\text{L}})\text{Bin}(n, p_{6\text{H}})\text{Bin}(n, p_{30}). \quad (7)$$

We use the same procedure for the eighth characteristic sequence pattern. When multiple items are tied for being the most frequently sampled item within a particular sequence, if any of those items are chosen, then in the empirical and simulation data we score it as an instance that the most frequently sampled item was chosen.

We use AIC to compare how well each decision rule matches the empirical sequence data while controlling for the number of parameters in a rule, and we use Akaike weights to estimate the probabilities that each decision rule best describes a given empirical data set. The CB and FT decision rules have four fitted parameters (K), while the BN rule has three. The AIC is two times the negative log likelihood plus a penalty for the number of fitted parameters:

$$\text{AIC} = -2\log\mathcal{L} + 2K. \quad (8)$$

A lower AIC means a better fit of the model with the observed data (Anderson et al. 2000). For each characteristic sequence pattern and for each decision rule, we identify the 10 best parameter combinations (i.e., the combinations that produced the smallest AICs). We then run these best parameter combinations again in simulations with 1,000 replicates. Since the simulations have stochastic components, we use a two-step process to avoid having a rule by luck produce a good fit with the observed data. Finally, for each characteristic sequence pattern and for each decision rule, we identify the best parameter combination (i.e., the combination that produced the smallest AIC). We use these exemplars of each decision rule to compare the relative likelihoods of the alternative decision rules at forming the empirical sequence patterns.

Decision rules are ranked by their AIC, and the rule with the lowest AIC is the best rule for approximating the information in the data (Anderson et al. 2000). It is the differences between AIC that are important, and as a general rule an AIC difference of more than eight might be considered a statistically significant difference (Burnham and Anderson 2002). The constant (c) in equation (3) and whether order is incorporated into calculations of probabilities both affect the magnitudes of AIC but do not affect the differences between AIC.

We convert AIC values into Akaike weights (Anderson et al. 2000). We rescale AIC values such that the lowest AIC has a value of 0:

$$\Delta_i = \text{AIC}_i - \min\text{AIC}. \quad (9)$$

The Akaike weight, which can be interpreted as the approximate probability that decision rule i is the Kullback-Leibler best model among the set of models under consideration (Anderson et al. 2000), is

$$w_i = \frac{\exp[-(1/2)\Delta_i]}{\sum_{r=1}^{R}\exp[-(1/2)\Delta_r]}, \quad (10)$$

with R being the number of alternative models.

Results of Likelihood Approach

For five of the seven characteristic patterns we evaluated, the CB rule produced the lowest AIC and, judging by the Akaike weights, was virtually certain to be the best decision rule for approximating the observed data (table 2). One exception was the pattern of returns, for which the FT rule produced the lowest AIC and, according to the Akaike weights, was virtually certain to be the best decision rule for approximating the data. The other exception was the frequency at which the last newly sampled item was chosen, for which the BN decision rule produced the lowest AIC and, according to the Akaike weights, was probably the best decision rule for approximating the data (table 2). The Akaike weights indicate that the approximate probability that the BN rule was the best rule was 79% versus 21% for the CB rule. For three of the characteristic seven patterns, the BN rule was shown to be incapable of producing the observed pattern because it does not produce enough variation to create the observed range of sequence patterns.

Table 2: Summary of Akaike Information Criteria (AIC) and Akaike weights (w) for the three decision rules when compared with patterns produced during western scrub-jay peanut sampling

	Comparative Bayes		Flexible threshold		Best of n	
Characteristic sampling sequence patterns	AIC	w	AIC	w	AIC	w
1. Items sampled	2,847.1	1	3,007.5	0		
2. Sample number	3,373.7	1	3,465.5	0		
4. Pr(return to immediately prior|return)	53.5	1	137.3	0	123.5	0
5. Pr(return), final and not final steps in sequence	434.0	0	404.1	1		
6. Pr(repeat), final and not final steps in sequence	1,020.9	1	1,266.2	0	1,701.3	0
7. Pr(chose last sampled)	22.6	1	75.0	0	229.4	0
8. Pr(chose most sampled)	58.8	.21	66.6	0	56.1	.79

Note: Scrub-jay data derived from Langen (1999). Empty cells indicate that the model never produced the characteristic empirical sampling pattern. Note that the third characteristic scrub-jay sampling sequence pattern is not evaluated, since it is not independent of the first two patterns.

Validation of the Likelihood Approach

To validate the performance of our model selection approach, we use simulations to create sampling sequences using each decision rule and then use the likelihood methodology to test whether this approach correctly identifies the decision rule that created the data. For each decision rule, we run 200 replicates of sampling sequences from six-item (high and low variance) and 30-item pools. We format and categorize the data from the simulations and apply the same likelihood approach as described in "Results of Likelihood Approach."

The decision rule that generated the simulated sequence data was inferred to be the best rule for approximating the data for most characteristic sequence patterns (tables 3–5). Therefore, our procedures and conclusions regarding the scrub-jay data are generally reliable. There were instances, however, in which an incorrect decision rule was indicated to be the best rule for approximating a particular characteristic pattern. This may result from stochasticity in the fitted data and the data produced to measure the fit.

Discussion

We find that the CB decision rule is the most likely of the three decision rules we have evaluated to produce the characteristic scrub-jay nut-sampling sequence patterns. The CB rule differs from the alternative rules in two principal ways. First, the CB rule allows jays to prospectively select which nut they will sample, whereas the FT and BN rules posit that the birds randomly encounter items. Second, the CB rule posits that jays use Bayesian updating to incorporate new information acquired by sampling to refine their estimate of the quality of an item. They track the quality of each item, and their estimates incorporate uncertainty about the quality of each. In comparison, the BN rule posits that jays remember the identity and quality of the best item they have encountered in a sequence, and the FT rule posits that jays ignore or forget about the quality of items sampled and rejected earlier within a sampling sequence.

Many models of choice behavior incorporate some form of Bayesian information assessment (Giraldeau 1997), and it appears to be taken for granted that animals use Bayesian updating processes to track the environment. In reality, there are few studies that provide empirical evidence that animals use Bayesian information assessment; most such studies focus on the dynamics of foraging in a patchy environment and evaluate a small number of straightforward but probably nonunique predictions generated by a Bayesian assessment model (e.g., Valone 1992; Wilhaber et al. 1994; Alonso et al. 1995; Killeen et al. 1996; Nonacs and Soriano 1998; Ollson et al. 1999; see also Hunte et al. 1985). Our study is unusual in that we evaluate a larger number of patterns in the empirical data and ask whether a Bayesian assessment model (or two alternatives) can account for them. The relative success of the CB decision rule model at reproducing the distinctive and nonobvious characteristic nut-sampling sequence patterns of western scrub-jays is, we believe, the most convincing evidence that exists at present that any animal uses a process analogous to Bayesian updating to modify information represented in its brain. Moreover, if western scrub-jays are indeed using an assessment process similar to the CB decision rule, then these birds are able to prospectively calculate the value of information gained by sampling particular items (sensu Stephens 1989) and selectively sample those items that are likely to provide the most useful information for Bayesian updating.

Examining why the CB decision rule provides a better model for the process western scrub-jays use to select a nut than the other decision rules and why for some char-

Table 3: Summary of Akaike Information Criteria (AIC) and Akaike weights (w) for the three decision rules when compared with patterns produced by simulations of the comparative Bayes decision rule

	Comparative Bayes		Flexible threshold		Best of n	
Characteristic sampling sequence patterns	AIC	w	AIC	w	AIC	w
1. Items sampled	2,218.5	1	2,256.2	0		
2. Sample number	2,550.2	1	2,708.1	0		
4. Pr(return to immediately prior|return)	39.8	1	56.4	0	51.2	0
5. Pr(return), final and not final steps in sequence	96.8	1	894.3	0		
6. Pr(repeat), final and not final steps in sequence	37.0	1	314.0	0	178.8	0
7. Pr(chose last sampled)	23.4	.52	23.6	.48	188.4	0
8. Pr(chose most sampled)	21.2	1	131.4	0	55.0	0

Note: Empty cells indicate that the model never produced the characteristic empirical sampling pattern. Note that the third characteristic scrub-jay sampling sequence pattern is not evaluated, since it is not independent of the first two patterns.

acteristic sampling sequence patterns it does not provide the best fit of the three may suggest which model assumptions are critical for describing the data and what model refinements are needed. Two characteristic sequence patterns that provide the most support for the CB decision rule over the other two rules indicate nonrandom movement: the likelihood of returning to a previously sampled nut is higher for the immediately prior item than one sampled earlier in a sequence, and an item is more likely to be repeatedly sampled at the end of a sequence than earlier within it. Three processes can produce nonrandom sampling sequences. First, past experiences may guide assessment decisions, as in the Bayesian updating used in the CB decision rule. This requires that jays recognize items and remember past sampling experiences with those items. In addition to processes like Bayesian updating forming the observed patterns, individuals remembering past encounters but tending to forget the details of older encounters may lead them to more often return to items sampled immediately prior. Second, before sampling, items might not appear equal in quality. For example, even from a distance, some peanuts may appear to be larger than others. If scrub-jays preferentially sample those putatively "high-quality" peanuts within a candidate pool, they would probably return and repeat more often than the random expectation; data in Langen and Gibson (1998) indicate that jays visually inspect nuts and preferentially sample larger items. This second process, however, does not account for why scrub-jays more often return to the peanut sampled immediately prior than to other previously sampled peanuts or why returns and repeats are more common than expected at the end of sampling sequences. The third process is that movement is random but spatially localized; scrub-jays could be more likely to sample adjacent peanuts. This process would produce the characteristic sample sequence pattern of more returns to the peanut sampled immediately prior than to other previously sampled peanuts, but it fails to account for the fact that many of the repeated resamples are to nonadjacent items (T. Langen, unpublished data), and resample probabilities increase disproportionately at the end of a sequence. Thus, the characteristic sampling sequence patterns indicate that the process leading to nonrandom movement patterns in the scrub-jay data is sampling decisions made on the basis of prior experience within a sequence, a key feature of the CB decision rule. Whether these sampling decisions are made using a process like Bayesian updating or is affected by proximate constraints like forgetting is yet to be resolved.

Two characteristic sampling patterns in the scrub-jay data are not best matched by the CB rule. Returns are more likely at the end of sampling sequences than earlier within them, and the CB decision rule is the only one to produce this pattern (table 1). However, for the likelihood approach, we measured the likelihood of the three decision rules producing the observed number of returns both at the end and throughout sampling sequences and found the FT rule provides a better match than the CB rule. This primarily occurs because the CB rule consistently produces too high a frequency of returns both at the end and throughout sampling sequences when sampling from a 30-item candidate pool. Thus, the failure of the CB rule might be partially explained by the metric we used in the likelihood analysis being a poor substitute for the original pattern. However, this result does show that when there are numerous items in the candidate pool, scrub-jays are either unable or unmotivated to return to previously sampled items as frequently as the CB rule would predict. Modifications of the CB decision rule that incorporate plausible cognitive constraints may better account for this pattern, including a limitation on the number of items that can by represented in working memory, a decreasing probability of relocating an item as more items are subsequently sampled (a form of retroactive interference), and

Table 4: Summary of Akaike Information Criteria (AIC) and Akaike weights (*w*) for the three decision rules when compared with patterns produced by simulations of the flexible threshold decision rule

	Comparative Bayes		Flexible threshold		Best of *n*	
Characteristic sampling sequence patterns	AIC	*w*	AIC	*w*	AIC	*w*
1. Items sampled	2,244.2	1	2,296.8	0		
2. Sample number	2,577.1	0	2,538.2	1		
4. Pr(return to immediately prior|return)	249.6	0	61.4	.27	59.3	.73
5. Pr(return), final and not final steps in sequence	939.2	0	124.8	1		
6. Pr(repeat), final and not final steps in sequence	252.7	0	43.8	1	55.6	0
7. Pr (chose last sampled)	24.3	1	55.8	0	139.1	0
8. Pr (chose most sampled)	402.6	0	26.3	1	49.3	0

Note: Empty cells indicate that the model never produced the characteristic empirical sampling pattern. Note that the third characteristic scrub-jay sampling sequence pattern is not evaluated, since it is not independent of the first two patterns.

increasing uncertainty about the assessed quality of an item with time since sampling or as more items are subsequently sampled, a form of "forgetting."

The other mismatch is that the frequency of choosing the most sampled item is better matched by the BN decision rule than the CB rule. For the six-item candidate pools, the CB decision rule is similar in this characteristic pattern to the empirical data. But for the 30-item pools, the CB decision rule results in more frequent selection of the most sampled item in a sequence than is observed in the empirical data, whereas the BN rule almost perfectly matches the observed frequency. When there are numerous items from which to choose, the CB decision rule results in repeated sampling of a "short list" of apparently high-quality options, selecting from within this subset. Incorporating cognitive constraints into the CB decision rule, as described in the previous paragraph, may clarify why the apparent mismatch between model and data occurs in this instance, as too may empirical sampling sequence data from a wider range of candidate pool sizes and distribution that were available in this study.

The CB rule produces a better fit with the scrub-jay sampling sequence data than the other decision rules for the frequency of repeats within a sequence but nevertheless produces a much lower repeat frequency than the empirical data. For example, in the six-item, high-variance candidate pool, scrub-jays repeated sampling an item for their final sample in 73% of the sequences, whereas terminal repeat samples occurred in only 10% of the sequences for the CB rule. This discrepancy too may be resolved by incorporating cognitive constraints into the model, including the relative advantages of different choreographies of repeat sampling for acquiring information (as shown in Honey and Bateson 1996).

We have demonstrated two statistical approaches for evaluating alternative cognitive models with empirical data. The first approach, based on frequentist statistics, suffers because the conclusions are too sensitive to sample sizes and the chosen level of statistical significance and because there is no quantification of how well alternative models fit the observed data. These two problems relate to how *P* values are interpreted. A *P* value reports the probability of the data given the null hypothesis is true but is often misinterpreted as the probability that the null hypothesis is true given the data (Anderson et al. 2000). The magnitude of the *P* value should not be interpreted as the strength of evidence either for the null or the alternative hypothesis. This mistake in interpretation of the *P* value is probably so commonly made because we want our data to provide conclusive evidence for or against hypotheses.

With a large enough number of replicates of a model simulation, we can expect that at least half of the statistical tests will result in a rejection of the null hypothesis in a direction consistent with the empirical sequence data. By focusing only on whether each decision rule model generates patterns that significantly deviate from the null hypothesis in the same direction as the empirical sampling sequence data, the first approach to model evaluation emphasizes qualitative fit and ignores quantitative fit to empirical data. Cases in which the null hypothesis is not rejected neither support nor weaken the plausibility that a decision rule can account for the empirical sequence data. From a Popperian falsificationist perspective, instances for which a pattern in the simulation data is statistically significantly different from a null hypothesis in the opposite direction of the empirical data provide the most critical information, obligating one to reject the decision rule as a plausible explanation for the empirical sequence data. However, it seems quite plausible that as more patterns are tested, the more likely a rejection will occur. This is unsatisfactory; the aim of a study such as ours is to evaluate which of a defined set of cognitive models best reproduces the empirical behavioral data, not

Table 5: Summary of Akaike Information Criteria (AIC) and Akaike weights (w) for the three decision rules when compared with patterns produced by simulations of the best of n decision rule

	Comparative Bayes		Flexible threshold		Best of n	
Characteristic sampling sequence patterns	AIC	w	AIC	w	AIC	w
1. Items sampled	2,745.3	0	2,397.3	0	1,432.2	1
2. Sample number	1,996.4	0	2,816.7	0	657.1	1
4. Pr(return to immediately prior|return)	175.5	0	62.7	.91	67.3	.09
5. Pr(return), final and not final steps in sequence	841.0	0	331.7	0	115.6	1
6. Pr(repeat), final and not final steps in sequence	170.8	0	70.1	0	44.2	1
7. Pr (chose last sampled)	604.3	0	117.6	0	21.7	1
8. Pr (chose most sampled)	81.9	0	275.4	0	22.3	1

Note: Empty cells indicate that the model never produced the characteristic empirical sampling pattern. Note that the third characteristic scrub-jay sampling sequence pattern is not evaluated, since it is not independent of the first two patterns.

to reject all cognitive models because they provide imperfect characterizations of empirical sampling sequences.

The second statistical approach based on likelihoods indicates that for most of the characteristic scrub-jay nut-sampling sequence patterns, the CB decision rule is the most likely of the three evaluated rules to reproduce the empirical data. The relative likelihoods of each decision rule reproducing the scrub-jay sampling sequence patterns can be interpreted as the magnitude of support the empirical data provide each rule; they are meaningful only in relation to the other decision rules under consideration. For most characteristic sequence patterns, the CB decision rule was the best rule for producing the observed patterns. Thus, we conclude that the CB decision rule provides a better cognitive model of western scrub-jay peanut selection than the FT and BN rules.

It is also possible to quantify the likelihood that a decision rule will concurrently produce several of the characteristic scrub-jay sampling sequence patterns. If the patterns are independent, one can simply sum the likelihoods of producing different behavioral patterns to find the likelihood of jointly producing all of the characteristic patterns. We have done this and found that the CB decision rule is far more likely to produce a combination of all of the patterns than the two other rules. Unfortunately, many of the patterns are not independent, and thus combining their likelihoods would be counting the evidence provided by the data more than once. However, by explicitly creating correlation matrices and weighting the evidence by those correlations, or by looking at the joint distributions of behavioral patterns, one may be able to combine the likelihoods of different metrics.

When there are alternative models of a process and simulations of these models can be used to create data in the same form as the empirical data, then a model comparison approach based on likelihoods is useful for evaluating the relative fit of each model to the empirical data. One can quantify the support for each of the alternative models without the traditional necessity of identifying and testing unique predictions of each. Beyond simply declaring which rule is the best match with the data, the likelihood approach facilitates examining where the models do and do not adequately describe the empirical data. For example, our comparison of the CB decision rule with the characteristic sampling sequence patterns indicates that it may be productive to evaluate the relative importance of spatial influences on sampling patterns versus the prospective value of information of sampling particular items as causes of nonrandom sampling sequence patterns, and we should incorporate cognitive constraints on assessment and evaluate how they affect the fit between model and empirical data. The approach facilitates a tighter link between models and data than the null hypothesis testing approach. Through an iterative process of designing models, gathering empirical data, comparing the data to the models, and updating the models, we can make greater progress at understanding cognitive processes underlying decision making and choice behavior and, indeed, many other complex biological processes.

Acknowledgments

The empirical and theoretical research that eventually led to this article was originally undertaken because of the inspiration and encouragement of R. Gibson, whom we warmly thank. The western scrub-jay data was collected (in part) at the University of California, Los Angeles, Stunt Ranch Santa Monica Mountains Reserve of the University of California Natural Reserve System, while T.A.L. was supported by a National Institutes of Health National Research Service Fellowship. The data analysis was conducted as part of the Evidence Project Working Group supported by the National Center for Ecological Analysis and Synthesis, a center funded by the National Science Foundation

(grant DEB-0072909), the University of California, and the Santa Barbara campus. Additional support was also provided for the postdoctoral associate B.L. in the group. We thank S. Lele, M. Mangel, M. Taper, M. Towner, D. Wiegmann, and an anonymous reviewer for their helpful comments on this article.

Literature Cited

Abrahams, M. V. 1986. Patch choice under perceptual constraints: a cause for departures from an ideal free distribution. Behavioral Ecology and Sociobiology 19:409–415.

Abrams, P. A. 1999. The adaptive dynamics of consumer choice. American Naturalist 153:83–97.

Alonso, J. C., J. A. Alonso, L. M. Bautista, and R. Muñoz-Pulido. 1995. Patch use in cranes: a field test of optimal foraging predictions. Animal Behaviour 49:1367–1379.

Anderson, D. R., K. P. Burnham, and W. L. Thompson. 2000. Null hypothesis testing: problems, prevalence, and an alternative. Journal of Wildlife Management 64:912–923.

Bensch, S., and D. Hasselquist. 1992. Evidence for active female choice in a polygynous warbler. Animal Behaviour 44:301–312.

Bertorelle, G., A. Biazza, and A. Marcaonato. 1997. Computer simulation suggests that the spatial distribution of males influences female visiting behaviour in the river bullhead. Ethology 103:999–1014.

Brown, L. 1981. Patterns of female choice in mottled sculpins (*Cottidae Teleostei*). Animal Behaviour 29:375–382.

Burnham, K. P., and D. R. Anderson. 2002. Model selection and multimodel inference: a practical information-theoretic approach. Springer, New York.

Chamberlin, T. 1890. The method of multiple working hypotheses. Science 15:92–96.

Clark, C. W., and M. Mangel. 2000. Dynamic state variable models in ecology. Oxford University Press, New York.

Curry, R., A. T. Peterson, and T. A. Langen. 2002. The western scrub-jay (*Aphelocoma californica*). Pages 1–36 *in* A. Poole and F. Gill, eds. The birds of North America. No. 712. Birds of North America, Philadelphia.

Dale, S., and T. Slagsvold. 1996. Mate choice on multiple cues, decision rules and sampling strategies in female pied flycatchers. Behaviour 133:903–944.

Dale, S., H. Rinden, and T. Slagsvold. 1992. Competition for a mate restricts mate search of female pied flycatchers. Behavioral Ecology and Sociobiology 30:165–176.

Fiske, P., and J. A. Kalas. 1995. Mate sampling and copulation behaviour of great snipe females. Animal Behaviour 49:209–219.

Gibson, R. M. 1996. Female choice in sage grouse: the roles of attraction and active comparison. Behavioral Ecology and Sociobiology 39:55–59.

Gibson, R. M., and T. A. Langen. 1996. How do animals choose their mates? Trends in Ecology & Evolution 11: 468–470.

Giraldeau, L. A. 1997. The ecology of information use. Pages 42–68 *in* R. Krebs and N. B. Davies, eds. Behavioral ecology: an evolutionary approach. 4th ed. Blackwell, London.

Gladstein, D. S., N. F. Carlin, and S. N. Austad. 1991. The need for sensitivity analyses of dynamic optimization models. Oikos 60:121–126.

Hilborn, R., and M. Mangel. 1997. The ecological detective. Princeton University Press, Princeton, N.J.

Honey, R. C., and P. Bateson. 1996. Stimulus comparison and perceptual learning: further evidence and evaluation from an imprinting procedure. Quarterly Journal of Experimental Psychology 49B:259–269.

Hunte, W., R. A. Myers, and R. W. Doyle. 1985. Bayesian mating decisions in an amphipod, *Gammarus lawrencianus* Bousfield. Animal Behaviour 33:366–372.

Janetos, A. C. 1980. Strategies of female mate choice: a theoretical analysis. Behavioral Ecology and Sociobiology 7:107–112.

Killeen, P. R., G.-M. Palombo, L. R. Gottlob, and J. Beam. 1996. Bayesian analysis of foraging by pigeons (*Columba livia*). Journal of Experimental Psychology Animal Behavior Processes 22:480–496.

Langen, T. A. 1999. How western scrub-jays (*Apelocoma californica*) select a nut: effects of the number of options, variation in nut size, and social competition among foragers. Animal Cognition 2:223–233.

Langen, T. A., and R. M. Gibson. 1998. Sampling and information acquisition by western scrub-jays, *Apelocoma californica.* Animal Behaviour 55:1245–1254.

Luttbeg, B. 1996. A comparative Bayes tactic for mate assessment and choice. Behavioral Ecology 7:451–460.

———. 2002. Assessing the robustness and optimality of alternative decision rules with varying assumptions. Animal Behavior 63:805–814.

———. In press. Female mate assessment and choice behavior affect the frequency of alternative male mating tactics. Behavioral Ecology.

McIntyre, N. E., and J. A. Wiens. 1999. How does habitat patch size affect animal movement? an experiment with darkling beetles. Ecology 80:2261–2270.

McKenna, C. J. 1985. Uncertainty and the labour market: recent developments in job-search theory. St. Martin's, New York.

Mitchell, R. 1975. The evolution of oviposition tactics in the bean weevil, *Calllosobrushus macalatus* (F.). Ecology 56:696–702.

Nonacs, P., and J. I. Soriano. 1998. Patch sampling be-

haviour and future foraging expectations in Argentine ants, *Linepithema humile*. Animal Behaviour 55:519–527.

Ollson, O., U. Wiktander, N. M. A. Holmgren, and S. G. Nilsson. 1999. Gaining information about Bayesian foragers through their behaviour. II. A field test with woodpeckers. Oikos 87:264–276.

Pitt, M. A., and I. J. Myung. 2002. When a good fit can be bad. Trends in Cognitive Sciences 6:421–425.

Platt, J. R. 1964. Strong inference. Science 146:347–353.

Real, L. 1990. Search theory and mate choice. I. Models of single-sex discrimination. American Naturalist 136: 376–405.

Reid, M. L., and J. A. Stamps. 1997. Female mate choice tactics in a resource-based mating system: field tests of alternative models. American Naturalist 150:98–121.

Rintamaki, P. T., R. V. Alatalo, J. Hoglund, and A. Lundberg. 1995. Mate sampling behaviour of black grouse females (*Tetrao tetrix*). Behavioral Ecology and Sociobiology 37:209–215.

Stephens, D. W. 1989. Variance and the value of information. American Naturalist 134:128–140.

Taper, M. L., and S. R. Lele. 2004. The nature of scientific evidence: statistical, philosophical, and empirical considerations. University of Chicago Press, Chicago.

Uy, A. C., G. Patricelli, and G. Borgia. 2001. Complex mate searching in the satin bowerbird *Ptilonorhynchus violaceus*. American Naturalist 158:530–542.

Valone, T. J. 1992. Information for patch assessment: a field investigation with black-chinned hummingbirds. Behavioral Ecology 3:211–222.

Valone, T. J., S. E. Nordell, L. A. Giraldeau, and J. Templeton. 1996. The empirical question of thresholds and mechanisms of mate choice. Evolutionary Ecology 10: 447–455.

Wiegmann, D. D., L. A. Real, T. A. Capone, and S. Ellner. 1996. Some distinguishing features of models of search behavior and mate choice. American Naturalist 147: 188–204.

Wilhaber, M. L., R. F. Green, and L. B. Crowder. 1994. Bluegills continuously update patch giving up times based on foraging experience. Animal Behaviour 47: 501–513.

Wittenberger, J. F. 1983. Tactics of mate choice. Pages 435–447 *in* P. Bateson, ed. Mate choice. Cambridge University Press, New York.

Associate Editor: Eldridge S. Adams