Biol Philos (2011) 26:419–437
DOI 10.1007/s10539-010-9205-7

REVIEW ESSAY

Evidentiary inference in evolutionary biology

Review of Elliott Sober's (2008) Evidence and evolution: the logic behind the science. Cambridge University Press, New York

James Justus

Received: 29 November 2009 / Accepted: 24 February 2010 / Published online: 16 March 2010

Introduction

The relationship between evidence and hypothesis provides the epistemic authority of empirical science. But standard statistical philosophies understand the relationship very differently, advancing incompatible inference methods that sometimes yield conflicting results about evidentiary support. These methodological tensions are manifested throughout evolutionary theory, from debates between Pearson and Fisher over biometry and chi-squared testing (Baird 1983; Morrison 2002), to recent controversies in phylogenetics about Bayesianism and model selection criteria (Ronquist 2004; Kelcher and Thomas 2007). This makes the clarity Sober brings to the subject scientifically valuable as well as philosophically illuminating. The wide variety of topics addressed and extensive utilization of technical work in statistics proper demonstrate mastery over the difficult issues involved.

Evidence and Evolution compiles and integrates many of Sober's recent publications. Chapter 1 outlines five methods of scientific inference—Bayesianism, likelihoodism, Fisher and Neyman-Pearson versions of frequentism, and Akaike model selection—and the assumptions and issues on which they depend, agree, and diverge. At 108 dense but accessible pages, it functions as an excellent detachable introduction to central topics in formal epistemology. This groundwork facilitates several precise and remarkably charitable formulations of the design argument in Chap. 2 that reveal its unredeemable flaws. Chapters 3 and 4 apply the same framework to fundamental issues in evolutionary biology: determining the relative significance of natural selection, drift, and phylogenetic inertia in producing

J. Justus (✉)
University of Sydney, Sydney, NSW, Australia
e-mail: jjustus@fsu.edu

J. Justus
Florida State University, Tallahassee, FL, USA

organismal traits; testing adaptive hypotheses with correlational and chronological data; assessing the proposed evidence for life's common ancestry; testing different hypotheses about phylogenetic relationships that assume common ancestry; and others. That *Evidence and Evolution* exhibits the same careful attention to detail, argumentative rigor, and philosophical depth as its published predecessors needs saying only once. Rather than blandly document this fact, I focus primarily on rare points of potential contention.

A formal epistemology primer

Disputes about the best basis for scientific inference have long exercised statisticians and philosophers, with biologists and other scientists occasionally joining the fray (e.g., Ellison 2004; McCarthy 2007). Chapter 1 appropriately begins with an overview of the concept of evidence and competing theories of inference useful to both biological and philosophical readerships. Conveying rudiments is the goal, not comprehensive exposition or resolution of longstanding disputes (with a few exceptions). Nor is a unified account of scientific inference presented. Sober's view samples across standard statistical philosophies depending on the context, but for principled reasons.

Following Royall (1997), Sober organizes the book around three important questions posed by evidence:

(1) What does the evidence say?
(2) What should you believe?
(3) What should you do?

Three prominent statistical philosophies are construed as answering each question.

According to Sober, **Bayesianism** answers (2) with a specific interpretation of Bayes' Theorem:

$$\text{(BT)} \quad p(h|o) = \frac{p(o|h)p(h)}{p(o)},$$

where h and o represent propositions, and $p(o) \neq 0$ is assumed. The relationship BT specifies is as uncontroversial as the standard Kolmogorov axioms that entail it. What is much more controversial is the Bayesian interpretation of the probabilities of BT as degrees of belief about a hypothesis (h) and observation(s) (o); the utilization of BT as a belief-updating mechanism; and, the claim BT provides the best basis for scientific inference. With a specific account of inference and belief change, Bayesianism constitutes a general epistemology of science.

To illustrate the view, consider Sober's example of testing the hypothesis that polar bears' current fur length is an adaptation (A). Suppose several bears are shaved or their fur is supplemented to different lengths. After the initial trauma, physiological properties positively correlated with the bears' fitness are then measured (e.g., body temperature, fertility, hunting efficiency, etc.) on the reasonable assumption that fur length is more likely an adaptation if fitness

depends strongly on it. These measurements are observations o that potentially bear on A. For Bayesians, determining the probability of A given the observations made, which Sober (8) somewhat unconventionally calls the *posterior probability* $p(\mathrm{A}|o)$, is the primary goal of scientific inference. This usually requires assessing the terms on the right hand side of BT. $p(\mathrm{A})$ represents the *prior probability* of A, i.e., the probability current polar bear fur length is an adaptation prior to observing the bears. $p(o)$ represents the unconditional prior probability of o, i.e., the probability the measurements would have the value they do before they are performed and independent of what A or other hypotheses say they should be. $p(o|\mathrm{A})$, called the *likelihood* of A, represents the probability of o given A, i.e., the probability the measurements would have the value they do *if* A were true. For the observations in this test, models of processes affecting polar bear fitness and/or their evolutionary history would help specify the likelihood. Developing the models upon which likelihood calculations are based is a primary focus of scientific research. Except perhaps for Bayesian scientists, estimating $p(o)$ and $p(\mathrm{A})$ are typically secondary.

These probabilities dictate required degrees of belief about the relevant scientific propositions for Bayesians. Suppose new evidence o is acquired, such as when the measurements described above are performed. Given the dependencies between A and o BT stipulates, this evidence might alter the degree of belief about A. A belief updating rule called *strict conditionalization* specifies how: the new probability of A after learning o, $p_{new}(\mathrm{A})$, is its former conditional probability on o, i.e., $p_{new}(\mathrm{A}) = p(\mathrm{A}|o)$.[1] In this way, Bayesians utilize BT to supply the evidentiary basis for scientific belief and answer (2).

Much more could be said about the basics of Bayesianism, which Sober says deftly, but this exposition motivates his main criticism: the seeming lack of evidentiary grounds for some of its key elements. Consider prior probabilities. Empirically well-confirmed models or frequency data *can* supply defensible priors. For instance, models of mammal evolution or data showing that other mammals' fur length is usually adaptive may provide strong evidence that current polar bear fur length is an adaptation, thereby justifying high priors. The problem is that priors of many scientific claims do not seem to have similar evidentiary bases, particularly for fundamental theories such as the general theory of relativity (GTR), or fundamental hypotheses such as that of life's common evolutionary ancestry (CA). Unlike the polar bear case, it is unclear what models or frequency data could be cited as sources of *evidentiary* support for $p(\mathrm{GTR})$ or $p(\mathrm{CA})$.

In principle, frequency data and/or well-confirmed models could also help in assessing $p(o)$, but similar problems occur. For example, Bayesians often invoke the fact that $p(o) = p(o|h)p(h) + p(o|\neg h)p(\neg h)$ to deflect the problem of estimating $p(o)$ to the relevant priors and likelihoods. But in addition to priors, evidentiary worries beset likelihoods such as $p(o|\neg h)$ where $h =$ GTR or CA. Such negative hypotheses, which merely deny well-studied, well-confirmed scientific hypotheses, are called *catchalls*. Again, the problem is the lack of relevant empirical

[1] Strict Bayesian conditionalization requires learning o with certainty, which is implausible for many types of scientific observations. *Jeffrey conditionalization* allows for belief updating with uncertain observation (Jeffrey 1983).

information—frequency data or highly confirmed models—by which to evaluate these probabilities.

But this criticism construes Bayesianism's broad tent—46656 distinct versions by Good's (1983) count—too narrowly. As long as BT's central role in updating belief and inference is respected, Bayesianism permits varying degrees of evidentiary constraint on prior probabilities. Some Bayesians, for example, require their degrees of belief conform strictly to what evidence warrants. Lewis' (1986) "Principal Principle" suggests one way of doing so; Pollock's (1990) principles of direct inference suggest another. Sober's criticism appropriately targets subjective Bayesians that allow degrees of belief to diverge from evidence, but is misplaced against those that do not.

Where subjective Bayesian ambitions outrun evidentiary considerations, Sober suggests the more modest question (1) is appropriate, and satisfactorily answered by the *law of likelihood*:

(LL) Observation(s) o supports (favors, confirms) h_1 over h_2 iff $p(o|h_1) > p(o|h_2)$, and $\dfrac{p(o|h_1)}{p(o|h_2)}$ designates the degree of support.

The inequality specifies a qualitative criterion and the ratio a quantitative measure of evidential support. **Likelihoodists** champion several features of LL:

(i) since scientific models typically supply the relevant likelihoods directly, evidential support is usually measured independently of prior and posterior probabilities;
(ii) relationships (ratios) between likelihoods, not their absolute values, determine degrees of support; and,
(iii) evidentiary support is *essentially* contrastive: whether o favors h_1 must be judged against other hypotheses (h_2 here) that also have definite likelihoods.

Although (i) and (iii) avoid reliance on empirically suspect priors and catchall likelihoods, Sober stresses their legitimacy when appropriately empirically grounded in frequency data or well-confirmed theory. As such, Sober (37) suggests likelihoodism is Bayesianism's tenable core, only diverging when Bayesianism problematically commits to empirically groundless probabilities:

> The likelihoodist is happy to assign probabilities to hypotheses when the assignment of values to priors and likelihoods can be justified by appeal to empirical information. Likelihoodism emerges as a statistical philosophy distinct from Bayesianism only when this is not possible.

A mathematical fact Sober (15) cites encourages this view: if $p(o|h) = p(o|\neg h)$, then $p(h|o) = p(h)$. That is, without a likelihood difference, observations are irrelevant to the posterior probabilities that are Bayesianism's primary focus. Likelihoods therefore represent the only conduit by which observations bear on hypotheses, and likelihoodism thereby seems to constitute Bayesianism's evidentiary basis.

This partial reconciliation is tempting, but somewhat overstated. Bayesians defend a measure of evidentiary support distinct from LL:

$$\text{(BS)} \quad \text{observation(s) } o \text{ supports } h \text{ iff } p(h|o) > p(h).$$

The inequality is equivalent to $p(o|h) > p(o|\neg h)$, and Sober notes that BS is a special case of LL's qualitative criterion where $h_2 = \neg h$. But unlike LL, BS is *not* essentially contrastive between h_1 and h_2 when $h_2 \neq \neg h_1$. Consider the bear measurements from above. To assess their empirical support for the hypothesis fur length is an adaptation (A) versus a product of pure drift (PD), LL counsels directly comparing their likelihoods with $\frac{p(o|h_1)}{p(o|h_2)}$. In contrast, Bayesians need two steps. First, the evidentiary support of o for A and PD must be individually assessed with whatever *quantitative* measure of Bayesian support satisfying BS is employed (see below). Then, the degree of evidentiary support in each case must be compared. Contrastive evidentiary support is therefore derivative on the non-contrastive concept for Bayesianism, whereas it is primitive for likelihoodism.

Besides this conceptual contrast, there are quantitative Bayesian support measures (which Sober [16] describes) incompatible with LL, such as the difference measure, $p(h|o) - p(h)$, and likelihood ratio measure, $\frac{p(o|h)}{p(o|\neg h)}$ (see also Fitelson 2007). When the priors and catchalls involved in these measures are empirically defensible, which likelihoodists demand, they often yield assessments of evidentiary support conflicting with LL. This, and whether evidentiary support is essentially contrastive, constitute fundamental differences between likelihoodism and Bayesianism, and show that Royall's (1997) questions inadequately distinguish these statistical philosophies. Bayesians are ambitious. They endeavor to determine what one should believe (2) *and* what evidence says (1). Fortunately, the cogency of Sober's evidence-based case for likelihoodism depends little on deciding which account of evidentiary support is best in these empirically tractable circumstances. But it does require a response to serious criticisms of the contrastive view of evidentiary support central to likelihoodism. It seems novel, competitor-less scientific theories are sometimes strongly confirmed simply because they account for previously unimaginable or poorly understood phenomena. The mass-energy relationship within relativistic physics is a candidate example (see Hellman 1999; Colyvan 1999). Such non-contrastive cases of evidentiary support constitute plausible counterexamples to likelihoodism. Sober leaves this important challenge unaddressed.

Sober's advocacy of likelihoodism ends with its general refusal to evaluate *composite* hypotheses. A *simple* hypothesis stipulates a single probability or probability distribution for whatever is being hypothesized about. For example, the hypothesis a coin is fair holds for exactly one probability, $p(\text{heads}) = 0.5$, and is therefore simple. *Composite* hypotheses, however, delimit families of such probabilities. The hypothesis a coin is biased, for example, holds for several probabilities of heads (all those $\neq 0.5$) and is therefore composite. If a composite hypothesis contains adjustable parameters—for example, if a linear relationship between two quantities is postulated but slope and intercept are unspecified—it is called a *model*. Models are therefore (infinite) disjunctions of simple hypotheses.

If tests like the polar bear experiment uncover fitness differences, biologists often believe this supports the composite hypothesis that natural selection produced the trait. The hypothesis is composite (and a model) because differences of *any* magnitude indicate the influence of selection pressure (disregarding measurement errors). The drift hypothesis, however, typically requires trait differences exhibit precisely 0 differential fitness, i.e., different fur lengths are equally fit.

The problem for likelihoodism is that composite hypotheses alone do not confer definite probabilities on specific observations, i.e., they do not provide likelihoods. Without further information, a coin being biased makes no probabilistic prediction about, say, five consecutive heads. The simple hypothesis $p(\text{heads}) = 0.5$ predicts it with precise probability $(0.5)^5$. For this reason, likelihoodists generally decline to evaluate models. But as the bear experiment and publications in evolutionary biology indicate, testing models is a core objective of the science. Likelihoodism's modesty is therefore excessive. Frequentists (and Bayesians) have no such qualms.

Frequentism unifies a motley collection of methods—Neyman-Pearson hypothesis testing, Fisherian significance testing, maximum likelihood estimation, Akaike model selection, etc.—around an epistemological view about sound scientific inference. Within Sober's question-based classification, these methods determine what should be done (question [3]) by focusing on their frequencies of favorable versus unfavorable outcomes when repeatedly applied (e.g., accurate vs. erroneous hypothesis evaluation). Sober describes many commonly cited difficulties with frequentist methods that cannot be recounted here: reliance on problematic formulations of probabilistic *modus tollens* (Royall 1997, Ch. 3); violation of the principle of total evidence (Howson and Urbach 1993, Ch. 8); dependence on (i) the (seemingly arbitrary) selection of what hypotheses are considered "null" and, (ii) what experimental stopping rule is in a researcher's head, which seems irrelevant to evidentiary considerations (Howson and Urbach 1993, Chap. 9).

Sober's assessment of frequentism turns from critical to favorable for a particular model selection method based on Akaike's theorem (Akaike 1973). Here's the idea (see also Kieseppä 2001). Consider datasets of paired points $\{(x_i, y_i) : i = 1, \ldots, n\}$ representing repeated sets of measurements of two quantities (e.g., fur length and fertility). The quantities are assumed to exhibit some (unknown) relationship. The objective is to order a set (**M**) of candidate models $\{M_1, \ldots, M_p\}$ based on their expected ability to predict future data generated by the underlying relationship. For example, two simple models of such a relationship are LIN, which postulates a linear relationship $(y = a + bx)$ involving two adjustable parameters (a and b), and PAR, which postulates a parabolic relationship $(y = d + ex + fx^2)$ involving three (d, e, and f).

An *estimator* of the predictive accuracy of a model M is *unbiased* if its expected value over indefinitely many datasets is M's true predictive accuracy. Given several assumptions (see below), Akaike proved:

(AT) $\ln\{p[\text{data}|L(M)]\} - k$ is an unbiased estimator of the predictive accuracy of model M.

$\ln\{p[\text{data}|L(M)]\} - k$ is called the Akaike information criterion (AIC) score for M. AT requires the datasets against which predictive accuracy is assessed contain the same number of datapoints. Formulating AIC as a per datum quantity circumvents this restriction (Sober 85).

AIC scores impose a ranking on **M**: higher scores indicate models' higher (estimated) predictive accuracy. The score has two components. k represents the number of adjustable parameters, so AIC's second term penalizes this type of model complexity. PAR, for instance, is penalized more than LIN. More parameters yield greater penalties, but also enhance the ability to fit data, which the first AIC term measures. $L(M)$ is the best fit member of M, called its *maximum likelihood estimate hypothesis*. Fitted members of M such as $L(M)$ are also sometimes called *curves* and AIC assesses a model's fit-to-data by how well its maximum likelihood curve fits the data, as gauged by the log-likelihood $\ln\{p[\text{data}|L(M)]\}$. In this way AIC balances fit-to-data against model complexity relative to the goal of maximizing expected predictive accuracy.[2] AIC favors LIN over PAR on simplicity grounds, but favors PAR over LIN on fit-to-data grounds. Their overall AIC score takes both desiderata into account and induces an ordering of LIN and PAR. Bayesians approach model evaluation very differently. They assess a model's fit not in terms of its maximum likelihood curve, but rather by the *average* fit of the curves comprising the model.

Akaike's theorem is impressive. The true underlying relationship required to assess predictive accuracy directly is usually unknown. But *seemingly* without any knowledge about that relationship, AT shows a model's *future* predictive accuracy can be estimated from its *current* fit-to-data and number of adjustable parameters. In this sense, Forster and Sober (1994) claim that AT makes predictive accuracy "epistemically accessible." In addition, partly by focusing on evaluating models rather than simple hypotheses, AIC avoids many difficulties with other frequentist methods, such as requiring one hypothesis be "null." Unlike some frequentist methods (e.g., Fisherian significance testing), AIC is also essentially contrastive: what matters is how AIC scores compare among candidate models in **M**, not their absolute values.

These advantages must be weighed against AT's limitations. Sober (87) describes three presuppositions of AT:

(i) repeated estimation of each model parameter forms a normal distribution (normality of measurement error);
(ii) the underlying unknown relationship being investigated does not change across datasets; and,
(iii) new data against which predictive accuracy is assessed are generated using the same overall process (experimental design, sampling procedure, etc.) from the same distribution of values as the original data. For example, if the initial fur length–fitness data was randomly sampled from just the 10–50 cm range, future data must also come from this range using random sampling.

[2] More precisely, AIC estimates the Kullback–Leibler discrepancy between $L(M)$ and the true curve representing the unknown relationship (see Burham and Anderson 2002, Chap. 2).

(i) and (ii) are relatively unproblematic assumptions made throughout statistics. (iii), however, restricts AIC to evaluating *interpolative* predictive accuracy (Forster 2000); *extrapolative* predictive accuracy falls outside AT's scope. This constitutes a serious limitation because extrapolation is as, or more, important to scientific inference than interpolation. Extrapolation underlies, for example, revered cases of significant scientific progress involving novel predictions. Myrvold and Harper (2002) argue Newton's derivation of a gravitational inverse-square law from information on planetary orbits is such an example. If these kinds of inferences permeate science and in fact facilitate its most important theoretical advances, as seems plausible, AIC fails to capture the indispensable nucleus of scientific inference.

This shortcoming prompts doubt about whether the balance between simplicity and fit-to-data AT brokers provides much support for general philosophical conclusions about science, such as that it "breathes new life into" (97) or, stronger, "makes plausible" (98), the epistemic instrumentalist view that "*the* goal of scientific inference is to find theories that make accurate predictions, not to find theories that are true" (Sober 97; emphasis added). *Perhaps* it can be agreed that the existence of a specific formal result about interpolative predictive accuracy such as AT somehow revitalizes, or even makes *more* plausible, instrumentalism about model-based scientific inference. But the restriction to interpolation, that there are many other estimators of predictive accuracy besides Kullback–Leibler discrepancy (Zucchini 2000), and the fact that numerous other inference methods are utilized within science warrants serious doubt about any stronger claim. As Sober notes, other model selection criteria that similarly balance model simplicity and fit-to-data have objectives other than predictive accuracy that are associated with non-instrumentalist views of scientific inference, such as scientific realists' attempt to maximize expected posterior probability with the Bayesian information criterion BIC (Schwartz 1978). Yet these criteria seem to serve the same function for non-instrumentalist philosophical views as AIC does for instrumentalism. Why AIC accrues more philosophical weight than other criteria is unclear.

Even when interpolative predictive accuracy is the goal, other limitations revealed within the cottage industry catalyzed by Forster and Sober's (1994) seminal paper deepen reservations about the AIC-based case for instrumentalism (e.g., Kieseppä 2001; Myrvold and Harper 2002; Dowe et al. 2007). For instance, that a model's true predictive accuracy is the *mean* of its estimated AIC scores over indefinitely many datasets (i.e., AIC is an *unbiased estimator*) is consistent with very large estimate *variance* (Sober 86–87). That is, estimates for particular datasets may differ significantly from true predictive accuracy. But slightly biased very low variance estimators provide more reliable inferences than high variance unbiased estimators for the *finite* numbers of datasets studied within science, so merely being unbiased does little to establish AIC is best among, or even competitive against, alternative inference methods (Myrvold and Harper 2002).

If AIC's variance were minimal or typically smaller than alternative methods for most important inference problems in science, this would buttress its support for instrumentalism about model-based inference. Although this remains an open question, several studies show AIC is inapplicable for a wide-variety of statistical

inference problems, more biased for finite numbers of datasets, or outperformed *on predictive accuracy* by other model selection criteria (see Boik 2004; Dowe et al. 2007). Part of the problem is that AIC is only an unbiased estimator *if* the true underlying relationship is a member of one of the models being analyzed. AT therefore requires some knowledge of the true relationship, specifically, that it is included in at least one of the models being considered. Sober (92) criticizes BIC for a similar assumption, but AIC is in the same boat. How the set of candidate models **M** should be selected, about which AT provides no guidance (Kieseppä 2001), is therefore crucial to AIC's performance as a predictive accuracy estimator.

Admittedly, a glaring liability besets many of these criticisms. They often rely on Bayesian model-selection criteria that require prior probability distributions over the infinite disjunctions of curves comprising the models analyzed. Sober disparages the empirical credentials of such distributions (see above) but the superior performance of these criteria in some contexts makes a tradeoff—rather than outright anti-Bayesian verdict—appear attractive: empirically suspect priors versus poorer predictive performance. If maximizing predictive accuracy is the primary objective of science, perhaps evaluating these priors and the posterior probabilities of hypotheses is necessary. The tradeoff is especially tempting if criteria other than AIC, such as BIC or the minimum message length criterion Dowe et al. (2007) analyze, were minimum variance estimators in a wide variety of inference problems, though Forster and Sober (forthcoming) argue AIC estimates interpolative predictive accuracy with lower expected error than BIC. The AIC-based case for instrumentalism critically depends on this issue, but it is far from resolved.

Testability and design

Having defended likelihoodism for simple hypotheses and AIC for models, Chap. 2 utilizes both to formulate and then dismantle various design arguments. Chapter 1 indicates why Bayesian formulations of such arguments are unhelpful: they require prior probabilities for which no empirical information or otherwise defensible evidentiary basis exists. Likelihoods provide a better framework.

Sober rejects two diagnoses of design arguments common among biologists:

(i) the observed adaptedness of organisms (O) makes the intelligent design (ID) hypothesis more likely than the hypothesis random chance (C) produced it, $p(O|ID) \gg p(O|C)$, but Darwinian evolutionary theory (DE) is much more likely than ID, $p(O|DE) \gg p(O|ID)$ (e.g., Coyne 2009);

(ii) the "no-designer-worth-his-salt" objection: apparently maladaptive or poorly adapted organismal features, such as the Panda's thumb or choke-prone structure of the human throat, demonstrate that $p(O|DE) \gg p(O|ID)$ (e.g., Gould 1980).

(i) concedes Paley's original design argument, pre Darwinian theory, was correct, which Sober rejects. That God would not create adaptively suboptimal traits is a suppressed premise of (ii) that creationists legitimately contest, typically by spinning some yarn about God's covert plan for these traits. Sober (128) rightly

stresses that *independent evidence* of a (the?) creator's goals and abilities is needed, rather than mere *suppositions* about designer psychology (see also Kitcher 1983, Ch. 2).

The traditional God concept is not Sober's target, but (ii) does seem effective against it. If God must be omnibenevolent, omniscient, and omnipotent, such an entity would deplore the undeserved disutility many maladaptive traits produce and would have the foresight and power to change things accordingly. It seems utterly implausible, for instance, that an omnibenevolent entity with the power and knowledge to prevent it could condone several hundred children annually choking to death or several thousand life-threatening abdominal pregnancies in the US alone (Atrash et al. 1987), easily avoided by restructuring throat physiology or bridging the small ovary–Fallopian tube gap (Coyne 2009). In effect, such traits license applying problem-of-evil considerations to questions of organismal design (contra Sober 167). If traditional theistic assumptions about God's nature are immutable, maladaptive traits do make *ID* highly unlikely.

Sober, like Hume, believes Paley's original design argument was flawed. Formulated with likelihoods, two valid, structurally similar design arguments focused on eyes and watches, respectively, reveal the difficulty (see Sober 141). For Sober, the similarity is misleading because both arguments require missing premises, only one of which is defensible. The watch argument's missing premise concerns the putative designer's desire-ability profile:

$(A_{favorable})$ if an intelligent designer made watches, it would have *wanted* (above all) and been *able* to give them features $G_1, G_2, \ldots G_n$;

where $G_1, G_2, \ldots, G_n$ are watch features taken to evidence design. $(A_{favorable})$ preserves the argument's soundness, but a different profile does not:

$(A_{unfavorable})$ if an intelligent designer made watches, it would have *wanted* (above all) and been able to *prevent* them having features $G_1, G_2, \ldots G_n$.

If $(A_{unfavorable})$ is true, the likelihood an intelligent designer created the watch is zero, so ensuring soundness requires a reason to reject $(A_{unfavorable})$. Fortunately for the watch argument, plenty of inductive evidence exists—e.g., intentions of actual human watchmakers and their known construction of watches—to support $(A_{favorable})$ and deny $(A_{unfavorable})$. Unfortunately for the eye design argument, nothing analogous ensures soundness. Assumptions about designer psychology often simply beg the existence question, as theological theorizing about God's nature based on biblical exegesis typically does. Unfailingly, such assumptions are empirically untestable: no non-question begging measurement procedure or otherwise defensible method for determining the designer's abilities and wishes is ever given. But absent independently warranted claims about creative intentions, i.e., claims justified without assuming a creator(s) exists, the eye design argument is unsound.

This kind of criticism has been developed elsewhere, but usually without the formal sophistication (and charity) Sober deploys. Its core is the untestability

charge. The charge is well-founded against the design argument, but the modality of 'test*able*' leaves room for philosophical stratagem, no matter how desperate and unmotivated. Logical empiricists attempted to circumvent this muddle by devising nonmodal formal testability criteria. This project has a long, complicated history, which has been summarily deemed a failure (Soames 2003). With an eye towards similarly ending modal maneuvering while avoiding past mistakes, Sober (151) tentatively proposes a new testability criterion (TC):

> Hypothesis H_1 can now be tested against hypothesis H_2 iff there exist true auxiliary assumptions A and an observation statement O such that (i) $p(O|H_1 \wedge A) \neq p(O|H_2 \wedge A)$, (ii) we now are justified in believing A, and (iii) the justification we now have for believing A does not depend on believing that H_1 is true or that H_2 is true and also does not depend on believing that O is true (or that it is false).

TC avoids the problems the logical empiricists previously encountered, but with a hefty cost: imprecision and blunted critical force. Sober (149) rightly emphasizes that logical empiricists attempted to formulate testability criteria with an impoverished set of tools, first-order deductive logic and the concept of an observation statement primarily. But their austerity was well-motivated. Without such precision, targets of the "untestable" criticism (metaphysicians for logical empiricists) could (and did) leverage its imprecision in their favor. One worries that some components of TC—'justification,' 'belief,' even 'depend'—afford ID "theorists" just such an opportunity. Without much more exact theories of these folk concepts than usually on offer, a project which Sober (152) sensibly avoids, TC seems to provide no additional ammunition against purportedly non-question begging, duly independent justifications for beliefs about a designer's properties.

Moreover, while championing AIC, Sober (81) defends maximization of predictive accuracy (rather than truth) as a legitimate scientific goal because it accords with the indisputable scientific practice of testing idealized (and therefore false) hypotheses. Investigating such hypotheses would be pointless if maximizing truth (i.e., posterior probability) were the only objective. But idealized *auxiliaries* are typically indispensable components of tests of idealized (and non-idealized) hypotheses. How this squares with TC's insistence that the auxiliaries be true is unclear. One might think Sober's notion of 'harmless idealizations'—false assumptions that nevertheless yield the same or very similar predictions as true assumptions when conjoined with hypotheses—provides the needed amendment. To know the relevant predictions agree, however, requires knowledge of what the true assumptions are (Sober 144, ff. 20), so the problem simply reemerges one step removed.

The temporal index 'now' introduces further maneuverability. Statements may fail TC because the independently attested auxiliary assumptions needed to evaluate them are unknown or technologically infeasible *now*. As knowledge grows, the known auxiliaries and technological sophistication required to assess them may change in currently unimaginable ways. Thus, untestability now may reflect in principle untestability, technological infeasibility of attesting the relevant auxiliaries, or limited imagination (Sober 149). If the second, the criticism is much too

weak. Various theories classified under the rubric 'string theory,' for example, are criticized for being currently technologically untestable (see Cartwright and Frigg 2007), but ID's shortcomings are surely more serious, and not just technological. If the third, the normative upshot of failing TC is unclear: ID "theorists" can always query the imaginative powers or epistemic capabilities of their detractors rather then accept the deficiency of in principle untestability. Different ways of being untestable now therefore have different evaluative implications for theories. And although a proposition being untestable now does help undercut claims it is known to be true, only *in principle* untestability seems to provide the normativity desired; the logical empiricists focused on it accordingly. Their testability criteria were not temporally indexed. Instead, they attempted to formalize (recursively) the idea that only observation statements or statements independently shown to be testable can defensibly serve as auxiliaries to demonstrate statements are in principle testable. The formal clarity of the criteria helped critics find their shortcomings, but also ensured their potential bite against metaphysics. Existentially quantifying over time as Sober (153) suggests does yield a conception of in principle testability, but not one that removes this problematic imprecision and the philosophical opportunity to leverage it.[3] As such, TC seems to discard the normative baby with the formal bathwater.

The rest of the chapter's analysis is unassailable. Sober shows how contemporary ID arguments appealing to the obscurantist notion of "irreducible complexity" exemplify the same deficiencies as Paley's, as well as garble core features of evolutionary theory such as the distinction between modeling infinite and finite populations, the sometimes non-gradual nature of evolutionary change, and the possibility of function switching. Sober also considers other design argument formulations, which fare similarly. These include an inductive argument relying on a veiled, highly problematic analogy that simply ignores evolutionary theory's explanation for adaptive organismal traits; and another that postulates different degrees of designer involvement in producing adaptive traits, but which AIC shows seriously falters on predictive accuracy. Chapter 2 concludes with an optimistic prediction. Current science suggests humans will likely be capable of constructing complex organisms from nonliving materials in the foreseeable future. When this occurs, Sober believes Paley's argument—which infers a deity exists from the existence of such organisms—will appear as bankrupt as other deluded but often quaint theistic caprices in human cognitive development. One can hope.

Testing selection, inferring ancestry

Rather than merely assert adaptive traits are due to natural selection or that all life has a common ancestor, which would be as vacuous as many ID "theories," biologists actually test evolutionary hypotheses. Chapters 3 and 4 consider

[3] The resulting notion is also too narrow. Technological, imaginative, or even budgetary limitations could contingently preclude statements from being testable at any time, but this reflects vicissitudes of human priorities and cognition rather than epistemic deficiency of the statements.

methodological issues these tests raise. Like Chap. 1, comprehensive exposition is not the goal, even if considerable ground is covered. Instead, simple but well-chosen biological examples drive the analysis.

Chapter 3 begins with an example intended to illustrate that testing even simple, highly idealized evolutionary models is surprisingly complicated and information intensive. Recall the polar bear experiment described above. Suppose measurements of fur length in a (finite) bear population form a distribution (e.g., normal distribution) in the centimeter range, say, [0,100] with a 10 cm mean. Call this mean trait value P (for present value). Two potential explanations of the evolutionary origin of such a trait are often considered: that pure-drift or selection-plus-drift produced it. Selection-plus-drift corresponds to the adaptation hypothesis A from above. Both explanations include drift because drift must occur in finite populations. Other mechanisms of evolutionary change exist, of course, and later in the chapter Sober (§3.10) considers the performance of specific models of selection-plus-drift versus phylogenetic inertia (see also Orzack and Sober 2001). An interesting and challenging continuation of Sober's analysis would explore niche construction's performance against selection-plus-drift as an explanation of adaptive organismal traits (see Sterelny 2005).

Biological drift and selection can be modeled in many ways. For the sake of illumination more than biological plausibility, Sober's focus is a very simple Brownian motion model of pure-drift. Call this model PD. PD claims 10 cm evolved as a random walk (e.g., Markov process) from some ancestral state, i.e., no length was favored in polar bear evolution because all lengths (excluding extremal cases) have identical fitness (see Hamilton 2009, Ch. 3). The specific model of selection-plus-drift Sober considers, call it SPD, claims the lengths are not equiprobable, in particular, that there is a fitness optimizing, probabilistic attractor fur length O responsible (at least partially) for P. To simplify the analysis, Sober assumes O is unchanging and unique; fitness decreases monotonically as fur length diverges from O; fur length fitnesses are frequency independent; and, he focuses strictly on *phenotypic* evolution (see Lande 1976).

The different fitness claims predict different population dynamics. If the ancestral state was a normal distribution with mean A, Sober (197) states that PD predicts it will flatten over time *while retaining this mean* (approaching a flat line as $t \rightarrow \infty$). In fact, the dynamics of even this simple model are more complicated. Sober (193) quite reasonably specifies that 0 and 100 cm are reflecting, rather than absorbing boundaries. This biologically realistic assumption stipulates that bears do not immediately expire or have their fur length permanently fixed if it reaches these lengths. Instead, unidirectional evolution away from these values towards 50 cm occurs. Although Sober is correct that PD predicts the population mean remains constant until individuals of the population reach the 0 and 100 cm boundaries, once they do the mean changes, evolving towards 50 cm as individual bears are reflected towards 50 cm (which is the mean when the distribution is a flat line at $t = \infty$). The mean stays fixed as $t \rightarrow \infty$ under PD only if $A = 50$ cm (cf. Sober 198, Fig. 3.4). Under PD, 50 cm effectively functions like a probabilistic attractor for mean fur length once random walks take some bears' fur length to 0 or 100 cm.

SPD predicts a similar, but less severe flattening caused by drift and a mean shift towards O due to selection (and potentially drift, depending on the details previously described). As Sober (197) explains, differential fitness of fur length (selection intensity), its heritability, and the effective population size determine the predicted shift velocity. As selection intensity, heritability, and the time between A and P increases, predictive differences between PD and SPD increase. For fixed heritability, evolution time, and effective population size note that PD and SPD predictions only differ to the degree fur length fitnesses differ. For small fitness differences, the hypotheses may be empirically difficult to distinguish.

Results of the fur-modification experiment help identify O if it exists. If bears with fur lengths near some specific value hunt more efficiently, are more fertile, better rear young, etc., then that value optimizes fitness. This experimental methodology does not assume the present mean fur length is fitness optimal, or utilize assumptions that presuppose PD or SPD hold; Sober later argues (208–209) that parsimony assumptions in common phylogenetic inference methods bias for PD. Based on such results, Sober believes the likelihoods of PD and SPD can sometimes be ordered. In particular, if P is "close enough" to O, this favors SPD over PD. Although subsequent sections provide helpful details about how A and O are estimated (§§3.3–3.5), how "closeness" is assessed for continuous traits such as fur length is unfortunately never specified beyond that it depends on the biological details mentioned above. Sober offers some guidance about the desired notion by distinguishing four ways a population can evolve from the ancestral state A to its current state P (indicated by the arrow '$\Rightarrow$' in cases [a]–[d]) in relation to the putative optimum O, and how they bear on PD and SPD. (from Sober 200, Fig. 3.6):

(a) $A \Rightarrow P = O$. Conclusion: SPD is more likely than PD.

The probability the mean fur length is *exactly* 10 cm is 0 on PD and SPD, so '=' here is approximate equality, not precise identity (Sober 192, ff. 4). Although the intended notion of '=' is not stated explicitly, presumably Sober thinks $P = O$ holds only if they are "close enough" in the sense indicated above. To avoid confusion, let '$=^*$' designate the intended notion.

(b) $P \Leftarrow A \quad O$. Conclusion: additional information required.

The population began in A and evolved *away from* the putative optimum O.

(c) $A \Rightarrow O \Rightarrow P$. Conclusion: additional information required.

The population evolved *through* O (the second arrow indicates that the population overshoots O, not that O itself changes).

(d) $A \Rightarrow P \quad O$. Conclusion: additional information required.

The population has evolved towards O.

Sober claims (a) establishes SPD is more likely than PD but that (b)–(d) are inconclusive without more information. The presentation of (a), however, is ambiguous. The issue is whether A, and that P has diverged from it, are known. The left side of (a) in Fig. 3.6 ($A \Rightarrow P$) seems to indicate that *yes* A is known, but the textual discussion is not as unequivocal. Sober mentions that arrows point from A to

P in each of (a)–(d) before analyzing them, but then only cites $P =^* O$ when discussing what inferences (a) licenses (see Sober 200, second paragraph). An earlier publication also supports a *no* answer, that A is unknown. After similarly construing (a) and then stating that (b)–(d) require additional information because $P \neq^* O$, Sober (2005, 135) clarifies that "If we can discover what the lineage's initial state [A] was, and this implies that (b) the population evolved *away* from the putative optimum, we're done–PD has the higher likelihood." But this implies (a) does not include knowledge about A. Sober (2005, Fig. 4) also glosses (a) as a 'yes' answer to the question: "Are P and O identical?" Thus, on the *no*-interpretation, Sober believes $P =^* O$ alone (rather than $A \Rightarrow P =^* O$) shows SPD is more likely than PD.

The interpretation matters. On the seemingly more accurate *no*-interpretation that A is unknown, the inference from (a) to SPD is mistaken for cases in which P is close enough to the putative O to warrant $P =^* O$ but where $P =^*$ 10cm is merely the byproduct of nonselective evolutionary drift from $A =^*$ 10cm. In such cases, $P =^* O$ masks the fact that $A =^* P$ and PD holds. With knowledge that $A \Rightarrow P$ and hence that $A \neq^* P$ on the *yes*-interpretation where A is known, the second interpretation presumably precludes these cases. But the inference also fails on the *yes*-interpretation when the evolution toward O is merely due to drift towards 50 cm, as explained above. If $O = 50$ cm for instance, population evolution towards O is consistent with both PD and SPD without further information. Thus, on both interpretations any inference (a) licenses depends on additional biological details besides that $A \Rightarrow P =^* O$ and is not reliable without them.

These clarifications reveal further difficulties with Sober's assessment that (a) is conclusive but (b)–(d) are inconclusive regarding PD versus SPD. Take (b). As Sober (201) recognizes, this case *seems* to support PD over SPD. If the hypothesized O must be probabilistically attractive and a population evolves away rather than towards it, this seems to establish O does not exist. Although Sober (2005) endorses this view (see above quote), Sober (201) now says it is "not always true," and he cites an example where $A = 10.1$, $O = 10.2$, and $P = 10$:

> [Suppose] the population has been evolving for a very long time. The lineage has evolved away from O, but it's still close. If there is only weak selection pushing the population towards 10.2 cm, it isn't surprising that it exhibits a trait value of 10 cm. On the other hand, if the PD hypothesis is true and the population evolves for a long time, the observed trait value of $P = 10$ cm is far less probable.

As the time between A and P increases, the disparity between what PD and SPD predict increases, but this may be insufficient to distinguish the hypotheses empirically if selective intensity is especially weak. For this reason, Sober says further details are needed to infer from (b) that PD is more likely than SPD. Agreed, but it seems (a) shares this shortcoming. Putting aside that *both* PD and SPD can explain a population's divergence from A in (a), knowledge that $A \Rightarrow P$ and hence that $A \neq^* P$ is necessary (but not sufficient) to infer that (a) favors SPD over PD. But just as the shift from $A = 10.1$ to $P = 10$ and away from $O = 10.2$ in (b) is misleading because P and A are "still close," it seems gauging $A \Rightarrow P$ in (a) raises

precisely the same worry. Without information on section intensity and evolution time, perhaps A and P are "still close" such that $P =^{*} O$ is inferentially unreliable, just as $A \Rightarrow P$ is misleading in Sober's specific (b) counterexample. Similar considerations hold for case (c) since it postulates the same evolution away from O at issue. In general, (b) and (c) seem to require no more or less biological details than (a) to license inferences about PD and SPD.

Part of what muddies Sober's analysis is lack of clarity about how A, O, and P are demarcated. Sober's discussion of population distributions, their means and shape, and how PD predicts they "squash" over time whereas SPD predicts less severe squashing and shifting towards O, suggests that standard statistical tests such as ANOVA would determine whether A, O, and P are sufficiently distinct or similar enough to warrant '$=^{*}$'. Sober (2005, 136) suggests precisely this, but it is absent from the book, perhaps because such tests of mean differences are typically frequentist, so Sober's previous criticisms would apply. Absent an alternative, assume this demarcation method. Now suppose $A = 11$, $O = 10$, and $P = 10.1$. Population size and the shape of the relevant distributions determine whether O and P (and A) are statistically significantly different, and thus whether these values correspond to (a) or (d). Selection intensity, assuming it exists, affects distribution shape. In particular, weak selection and small population size may produce distributions quite similar to those PD predicts, whereas strong selection and large population size may produce very different distributions. Given values for A, O, and P, determining which case [(a)–(d)] is represented therefore depends on further biological details, population size and selective intensity (and statistical assumptions) at least. Thus, even if (a) were inferentially privileged, to know it and not other cases are applicable requires the same biological details Sober pinpoints as the shortcoming of inferences based on (b)–(d) alone.

Besides these difficulties, there is an unresolved puzzle about this approach to evaluating PD and SPD. These hypotheses are about evolutionary history. The historical focus poses additional inferential challenges unaddressed by the distinctions between (a) and (d). As Sober carefully explains with examples drawn from evolutionary biology, *present* data on fur length optimality may mislead about *past* evolutionary forces responsible for current trait values (§3.3). For example, the environment could have changed and O with it. Moreover, populations have many evolutionary ancestors. If their ancestral mean fur lengths or past optimal fur lengths differed, then PD and SPD must be evaluated for each one. Apart from the problems confronting inferences based on (a)–(d), Sober emphasizes that these complications reveal limitations about what inferences the polar bear experiment justifies. But assume for the sake of argument that these complications are unproblematic; perhaps good climatic and paleontological data shows that present observations provide foolproof information about past trait fitnesses. Suppose the experiment is then performed and fitness differences are found. *Ex hypothesi* these differences correspond to past fitness differences. But given this, a conclusion seems to follow immediately *regardless* of the different relationships between A, O, and P described in (a)–(d): SPD is more likely than PD because PD is incompatible with any fur length fitness differences. That is, once the existence of fitness differences is established (i.e., an O is found), it seems we are done. Since distinguishing between

(a) and (d) provides no information about the relationship between past and present evolutionary forces—as suitable climatic or paleontological data does—the distinctions therefore seem superfluous to the issue of whether SPD or PD is more likely given that simply determining whether an *O* exists suffices. Although distinguishing (a)–(d) was intended to illuminate when SPD is more likely than PD or vice versa, the distinctions are in fact irrelevant to this issue. Instead, the research focus should be acquiring the relevant climatic and paleontological data. Note that establishing fitness differences affects the likelihoods of PD and SPD but does not *presuppose* either hypothesis. It thereby satisfies Sober's emphasis that assumptions used to test PD versus SPD must be independently *justified* (see Sober 145, 202-203). Sober sometimes characterizes the independence required as probabilistic independence (e.g., final paragraph of §3.10). Much of the analysis throughout the book would prove problematic with this sense of independence.

These difficulties are isolated to the initial sections of Chap. 3. After analyzing several more specific evolutionary models in later sections—e.g., when fitness declines nonmontonically with divergence from *O* or when traits are dichotomous rather than continuous—Sober considers a different method for testing PD and SPD common in evolutionary biology. It focuses on whether correlations between environmental factors and phenotypic traits *across species*, such as fur length and ambient temperature, favor SPD over PD (see also Millstein 2008). This circumvents many of the problems raised above by changing the explanandum. In particular, the method does not require information about *A*. SPD seems to predict such correlations if fur length influences fitness relative to ambient temperature, and would thus be favored over PD if they were found. Sober describes the biological details required to make this intuitive reasoning rigorous. Following an interesting detour about Reichenbach's principle of common cause (§3.8, reutilized in Chap. 4, §4.4), Chap. 3 concludes by arguing AIC offers a better methodology than Bayesianism, likelihoodism, or Neyman-Pearson frequentism for testing SPD versus PD about molecular evolution, and selection versus phylogenetic inertia (Wilson 1975; Orzack and Sober 2001) as competing accounts of phenotypic traits.

Chapter 4 considers when similarity across organisms provides evidence of common ancestry and, in general, a last universal common ancestor. Unsurprisingly, this inference's soundness depends on numerous biological details: how traits are individuated, measured (e.g., dichotomous, discrete, or continuous), and their similarity assessed; how exactly organisms reproduce; what evolutionary forces likely shaped the lineages being analyzed; degree of lateral gene transfer; whether traits influence speciation probabilities, etc. Despite this complexity, Sober shows that different types of similarities rank differently in evidentiary significance for a range of cases. Specifically, deleterious trait similarities provide stronger evidence of common ancestry than neutral traits, which provide stronger evidence than adaptive similarities. This analysis, and Sober's explanation of how AIC helps evaluate competing phylogenetic trees later in the chapter, is particularly insightful and compelling. It skillfully utilizes examples from evolutionary biology to reveal what details are needed to infer evolutionary relationships, and why. As in Chaps. 2–3, actual scientific methodology illustrates and helps support the view of evidentiary inference developed in Chap. 1.

Conclusion

Evidence and Evolution is a wide-ranging, technically sophisticated, and careful exploration of the philosophically rich intersection of formal epistemology and philosophy of biology. Since the work's cogency is obvious to anyone who reads the book carefully, this review has tried to focus on potential difficulties. Such difficulties are usually easy to find in philosophy. They were not in this case. Sober's new book is highly recommended.

Acknowledgments Thanks to Mark Colyvan, Jenann Ismael, Adam LaCaze, Katie Steele, and especially Aidan Lyon and Elliott Sober for helpful feedback. Reading group discussions at the University of Sydney and Florida State University were also beneficial. I am grateful to the Australian Commonwealth Environment Research Facilities Research Hub: Applied Environmental Decision Analysis and the Sydney Centre for the Foundations of Science for research support.

References

Akaike H (1973) Information theory as an extension of the maximum likelihood principle. In: Petrov B, Csaki F (eds) Second international symposium on information theory. Akademiai Kiado, Budapest, pp 267–281

Atrash H, Friede A, Hogue C (1987) Abdominal pregnancy in the United States: frequency and maternal mortality. Obstet Gynecol 69:333–337

Baird C (1983) The Fisher/Pearson chi-squared controversy: a turning point for inductive inference. Br J Philos Sci 34:105–118

Boik R (2004) Commentary. In: Taper M, Lele S (eds) The nature of scientific evidence. University of Chicago Press, Chicago, pp 167–180

Burham KP, Anderson DR (2002) Model selection and multimodal inference: a practical information-theoretic approach. Springer, New York

Cartwright N, Frigg R (2007) String theory under scrutiny. Phys World 3:14–15

Colyvan M (1999) Contrastive empiricism and indispensability. Erkenntnis 51:323–332

Coyne J (2009) Why evolution is true. Viking Adult, New York

Dowe DL, Gardner S, Oppy G (2007) Bayes not bust! Why simplicity is no problem for Bayesians. Br J Philos Sci 58:709–754

Ellison AM (2004) Bayesian inference in ecology. Ecol Lett 7:509–520

Fitelson B (2007) Likelihoodism, Bayesianism, and relational confirmation. Synthese 156:473–489

Forster M (2000) Key concepts in model selection: performance and generalizability. J Math Psychol 44:205–231

Forster M, Sober E (1994) How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions. Br J Philos Sci 45:1–35

Forster M, Sober E (forthcoming) AIC scores as evidence—a Bayesian interpretation. In: Forster M, Bandyopadhyay P (eds) The philosophy of statistics. Kluwer, Dordrecht

Good I (1983) 46656 Varieties of Bayesianism. In: Good thinking: the foundations of probability and its applications. University of Minnesota Press, Minnesota

Gould S (1980) The panda's thumb. Norton, New York

Hamilton M (2009) Population genetics. Wiley-Blackwell, New York

Hellman G (1999) Some ins and outs of indispensability: a model-structuralist approach. In: Cantini A, Casari E, Minari P (eds) Logic and foundations of mathematics. Kluwer, Dordrecht, pp 25–39

Howson C, Urbach P (1993) Statistical reasoning: a Bayesian approach. Open Court, Peru

Jeffrey R (1983) The logic of decision, 2nd edn. University of Chicago Press, Chicago

Kelcher SA, Thomas MA (2007) Model use in phylogenetics: nine key questions. Trends Ecol Evol 22:87–94

Kieseppä IA (2001) Statistical model selection criteria and the philosophical problem of underdetermination. Br J Philos Sci 52:761–794

Kitcher P (1983) Abusing science: the case against creationism. MIT Press, Cambridge
Lande R (1976) Natural selection and random genetic drift in phenotypic evolution. Evolution 30:314–334
Lewis D (1986) A subjectivist's guide to objective chance. In: Philosophical papers, vol II. Oxford University Press, Oxford
McCarthy MM (2007) Bayesian methods for ecology. Cambridge University Press, Cambridge
Millstein R (2008) Distinguishing drift and selection empirically: 'The great snail debate' of the 1950s. J Hist Biol 41:339–367
Morrison M (2002) Modelling populations: Pearson and Fisher on Mendelism and biometry. Br J Philos Sci 53:39–68
Myrvold W, Harper B (2002) Model selection, simplicity, and scientific inference. Phil Sci 69:S135–S149
Orzack S, Sober E (2001) Adaptation, phylogenetic inertia, and the method of controlled comparisons. In: Adaptationism and optimality. Cambridge University Press, New York
Pollock J (1990) Nomic probability and the foundations of induction. Oxford University Press, Oxford
Ronquist F (2004) Bayesian inference of character evolution. Trends Ecol Evol 19:475–481
Royall R (1997) Statistical evidence: a likelihood paradigm. Chapman & Hall, London
Schwartz G (1978) Estimating the dimension of a model. Ann Stat 6:461–465
Soames S (2003) The rise and fall of the empiricist criterion of meaning. In: Philosophical analysis in the twentieth century. Princeton University Press, Princeton, pp 271–299
Sober E (2005) Is drift a serious alternative to natural selection as an explanation of complex adaptive traits? R Inst Phil Suppl 80(56):125–153
Sterelny K (2005) Made by each other: organisms and their environment. Biol Phil 20:21–36
Wilson EO (1975) Sociobiology: the new synthesis. Harvard University Press, Cambridge
Zucchini W (2000) An introduction to model selection. J Math Psychol 44:41–61