

# *Spandrels* and a Pervasive Problem of Evidence

Patrick Forber  
Tufts University

March 7, 2009

## Abstract

Evolutionary biology, indeed any science that attempts to reconstruct prehistory, faces practical limitations on available data. These limitations create the problem of *contrast failure*: specific observations may fail to discriminate between rival evolutionary hypotheses. Assessing the risk of contrast failure provides a way to evaluate testing protocols in evolutionary science. Here I will argue that part of the methodological critique in the *Spandrels* paper involves diagnosing contrast failure problems. I then distinguish the problem of contrast failure from the more familiar philosophical problem of underdetermination, and demonstrate how contrast failure arises in scientific practice with an investigation into Lewontin and White's (1960) estimation of an adaptive landscape.

**Keywords:** adaptation, adaptive landscapes, confirmation, evolutionary biology, spandrels, underdetermination

## 1 Introduction

The *Spandrels* paper (Gould & Lewontin, 1979) raises a general problem of evidence for adaptationist inquiry: limitations on epistemic access to history detrimentally compromise testing methods.<sup>1</sup> Adaptationist hypotheses presume differences in fitness and such differences supervene

on ecology. In principle, we want evidence for the presence of the relevant ecological factors over evolutionary time. In practice, we lack access to the complete ecological history. Past conditions are inferred based on information about present ecology, and there is no assurance that the past environment resembles the present. *Prima facie*, the *Spandrels* paper exemplifies a pessimistic reaction to this problem, taking the lack of access to historical data to cause insuperable problems for adaptationist inquiry. In fact, the access problem affects all aspects of evolutionary inquiry. Phylogenetic hypotheses about evolutionary relationships and ancestral character states are tested by comparing extant taxa and the occasional fossil, and process hypotheses about molecular evolution are tested by examining patterns of variation in current populations. If we cannot obtain the data we want in principle, how can we rigorously test hypotheses about evolution in practice?

Disputes about method, found in the *Spandrels* paper in particular and throughout evolutionary biology generally, reveal both the importance and difficulty of constructing effective tests. The *Spandrels* paper accuses the adaptationist program of adopting a testing protocol that systematically fails to provide adequate evidence for hypotheses of adaptation. The primary problem is that the protocol neglects hypotheses about non-adaptive processes. Only adaptation hypotheses are considered and so evidence that favors these hypotheses over those that posit alternative processes is not sought. Yet hypotheses appealing to developmental constraint, neutral evolution, and indirect selection (on a linked trait or for some other function) are also capable of explaining the apparent optimality in the current state of a lineage. Phenotypic plasticity, for example, can generate the appearance of optimality without a history of direct selection (Gould & Lewontin, 1979, 592). This undermines the strategy of identifying optimal traits in the current environment as evidential support for historical hypotheses about natural selection: "One must not confuse the fact that a structure is used in some way...with the primary evolutionary reason for its existence and conformation" (Gould & Lewontin, 1979, 587). Sufficient evidence for an adaptation hypothesis must favor it over the relevant rivals. If we do not explicitly include those rivals in the testing protocol then we cannot assess whether available data meet this standard.

The problem is exacerbated by the complexity of the evolutionary process and the concomitant flexibility of our models. Our models may incorporate a number of biological parameters such that hypotheses empha-

sizing either selection, drift, or constraint can be contrived to fit almost any evolutionary pattern. The *Spandrels* paper claims that the adaptationist program exploits this to their advantage by telling “just so” adaptive stories, and inventing new adaptive stories in response to any evidence against adaptation (Gould & Lewontin, 1979, 586). Beatty (1987, 70) also emphasizes that selection hypotheses can be constructed to predict any particular evolutionary outcome and that this complicates tests of selection against drift. Sober (1996, 2005) points out that the danger of “just so” story telling applies to anti-adaptationist alternatives as well. Because the evolutionary process is complex and stochastic, and because we lack complete epistemic access to this historical process, the *Spandrels* problem is pervasive.

Here I want to diagnose this problem from the *Spandrels* paper as the problem of *contrast failure*, and argue that there is a clear way to resist pessimistic worries. Yet the *Spandrels* paper is right to insist that we should carefully ascertain the effectiveness of testing protocols. Assessing the risk of contrast failure—diagnosing situations where specific data may fail to discriminate between rival evolutionary hypotheses—is one way to carry out this evaluation. The source of contrast failure traces back to general limitations on epistemic access in evolutionary inquiry. So I will begin by tracing the problem to its source (§ 2). Once identified, I will show that the problem of contrast failure is a distinct strain of underdetermination (§ 3). I will then make clear how to assess the risk of contrast failure in scientific practice, and how to use the problem to evaluate actual tests, with the analysis of an illustrative case study: Lewontin and White’s (1960) classic estimation of an adaptive landscape for *M. scurra* (§ 4).

## 2 The problem

Evidence is *contrastive*. For some data to count as evidence they must favor one hypothesis over some set of alternatives.<sup>2</sup> Yet data may *fail* to provide evidence because they cannot discriminate between rival hypotheses. This is the problem of *contrast failure*. A specific contrast failure problem is defined relative to a set of rivals and a set of data or observations. If those observations fail to discriminate between the rivals then we have contrast failure. Data that favor one hypothesis over other rivals avoid contrast failure, and count as evidence for that hypothesis. Unpacking the

exact nature of the favoring relation is the business of formal confirmation theory. Although different theories of confirmation will deliver different fine-grained assessments of evidential favoring (Fitelson, 2007), all strive to capture the coarse-grained judgments of favoring taken from exemplary scientific practice. All theories of confirmation will have to reckon with the problem of contrast failure. Thus, investigating how limitations on epistemic access generate contrast failure problems lies beyond the scope of any specific confirmation theory.

And we clearly have limitations on our epistemic access to biological evolution. Relatively few remnants of the historical process remain. Biologists know that they lack access to events over deep evolutionary time, yet they attempt, and often succeed, to use what data they have to discriminate between rival hypotheses about pattern and process. Despite the individual successes, however, there is a broad pessimistic objection to current methods of evolutionary inquiry. Much like an intoxicated friend looking for her keys under the lamppost, biologists may, in practice, persist in looking for answers “where the light is”: evolutionary inquiry searches for answers where data can be found, even though such data fail to address the historical questions under investigation. The *Spandrels* critique accuses adaptationists of acting like our friend under the lamppost, taking current utility (the accessible data) to count as clear evidence for adaptation (a historical hypothesis). That is, the adaptationist program makes the wrong compromise in the face of limitations on epistemic access.

Consider a concrete case. In principle, testing whether selection or drift explains the evolution of crypsis (say) requires gathering data on factors such as predation pressures, the availability of suitable hiding spots in the habitat, and population sizes. In practice, we lack access to the ecological and demographic history of a population. Instead of directly investigating the history, we test whether the trait frequency of cryptic coloration correlates with an environmental variable, such as the availability of hiding spots, in the current generation. Suppose the selection hypothesis predicts a strong positive correlation whereas the drift hypothesis predicts no degree of correlation. These are statistical predictions; any degree of correlation is *consistent* with both hypotheses. Statistical predictions in evolutionary biology raise familiar problems about confirmation: which hypothesis do the data support and to what degree? Suppose a strong positive correlation obtains. Although the correlation test may lack the

rigor of a direct test of the historical commitments—the result is consistent with both hypotheses—the positive correlation favors selection over drift. Alternatively, suppose there is only a slight positive correlation. Now the issue of evidential support is not so clear, and is compounded by observational error and statistical confidence intervals. The risk of contrast failure increases. A similarly high risk is present in testing situations where hypotheses about evolutionary process make similar or even identical statistical predictions about an evolutionary outcome.<sup>3</sup> Overcoming contrast failure in this case requires additional data from (say) comparative methods or historical investigation.

This points to a connected problem. The correlation test provides evidence for only one of the many elements of adaptation hypotheses. A correlation test provides evidence that selection occurred, but not for the ecological, genetic, demographic, or phylogenetic elements of potential adaptation explanations, evidence that is necessary to complete the explanation (Brandon, 1990, 165). Meeting these evidential demands is difficult. Different commitments about these separate elements can generate distinct adaptation hypotheses that make the same statistical predictions, undermining the evidential import of correlation data. Perhaps crypsis evolved to fool territorial conspecifics rather than predators, or perhaps crypsis is linked to a different trait under different selection pressures. If confounding factors occur, then a positive correlation between cryptic coloring and the microenvironmental choices of the organism can fail to provide evidence for the hypothesis that crypsis is an adaptation for predator evasion.

Is this a case of simply looking where the light is? Given the risk of contrast failure, refining the question about the evolution of crypsis to work with available correlation data is bad scientific practice, or so the pessimistic objection goes. But the objection misses a crucial difference. Our friend's behavior is irrational if she is insensitive to background evidence that the keys are not to be found under the lamppost. Evolutionary biologists, on the other hand, should look where the light is, so long as they do not share this insensitivity. If there is a chance that the keys may be under the lamppost, and we have no background evidence to the contrary, then looking there first makes good sense. It is *good* scientific practice to refine questions in a way that permits available data to discriminate between them. The correlation test for selection can be derailed because such data may fail to favor one hypothesis over others. If the data fail to

contrast the hypotheses, scientists must look elsewhere.

Lewontin (2002) poses what I take to be another version of the contrast failure problem, and claims that this problem persists even with rich data sets. In response to the fluctuating selection pressures uncovered by the Grants' (1986; 1989; 2002; 2006) long-term study of Darwin's finches, Lewontin claims that:

The consequence of the weakness of selective and random forces is that the processes of evolution in living species cannot, except very rarely, be followed as a *dynamic* process in time. Instead, the evolutionary biologist must depend on *static* data, observations of patterns of variation within and between species, to infer the dynamic processes that could not be directly observed (2002, 3).

Lewontin diagnoses the problem with a distinction between the *dynamic* process and *static* data. In principle we want information about the dynamic process. In practice we have only static data: patterns of variation. These within and between species patterns of variation can be phenotypic, such as the varying beak sizes of individual finches, or genotypic, such as sequence differences in shared proteins, genes, or neutral regions of DNA. What makes these patterns *static* is that the data come from a single evolutionary time step, usually the current generation. Formal evolutionary models guide the inferences about dynamic processes by specifying the patterns indicative of the different processes. Sober (2008) illustrates how this process works with his abstract test of selection versus drift, a test that will be discussed at length below (§ 3). This sort of inference dominates the field of molecular evolution, in particular

If static data are often insufficient to follow the dynamic evolutionary process then what would count as sufficient data? That is, how can we obtain *dynamic data*? In principle, the maximal data set includes the phenotypic and genotypic characteristics of all individuals in the relevant lineages and their lines of inheritance across some scope of evolutionary time. Notice that Lewontin does not see the task of obtaining dynamic data as impossible, just extremely difficult. Thus, the maximal data set must count as sufficient for dynamic data. This contrast between static and dynamic data identifies a spectrum for evaluating the epistemic quality of a data set. If static data are patterns of variation from a single evolutionary

time step then dynamic data must be patterns taken from enough evolutionary time steps to sufficiently track the evolutionary process. How to set the threshold for sufficiency is a hard problem, and one that will be sensitive to the details of specific cases. Isolating the spectrum of epistemic quality Lewontin has in mind at least clarifies how we should approach this problem.

Further support of this interpretation of the difference between dynamic and static data comes a little later when Lewontin distinguishes between two types of inquiry:

Both the detailed study of particular natural historical cases of observed dynamical changes and the use of static data to infer unobservable dynamical forces have dealt with a small number of specific examples of general phenomena: How are changes in bill and body size in Darwin's finches to be explained by the observed reproductive behavior of finches? Is there evidence that amino acid replacements in alcohol dehydrogenase that occurred in the evolutionary divergence of two species of *Drosophila* were the result of natural selection? (2002, 6)

Lewontin here counts the Grants' study as an example of "observed dynamical changes" and so different and superior in epistemic quality to the sort of inference from static data found in molecular evolution. The latter kind of inference "uses static data to infer unobservable dynamical forces," exemplified by the referenced MacDonald & Kreitman (1991) study.

The problem, then, is that we usually lack access to the information sufficient to reconstruct the dynamic process over evolutionary time. We instead use data taken from one evolutionary time step, the current generation, and attempt to infer past dynamics. But *different* dynamic processes can produce the *same* patterns in the static data. Hence Lewontin complains that static data are generally inadequate for confirming hypotheses about evolutionary dynamics because such data fail to discriminate between different hypotheses about the process. Static data almost unavoidably faces a contrast failure problem with respect to rival hypotheses on evolutionary dynamics.

Lloyd's (1988) discussion of confirmation in evolutionary biology coheres with Lewontin's diagnosis of static data. Lloyd explicitly offers her account as a descriptive one (1988, 145), and argues that confirmation of evolutionary models, specifically population genetic models, comes in

three ways (1988, 146–152): fit between model predictions and data, independent tests of assumptions of the model, and variety of evidence. Lloyd presents several case studies to show that in-practice testing involves more than assessing fit between predicted and observed evolutionary outcomes. Observed evolutionary outcomes are static data: the resulting patterns of variation found in the current generation. Lloyd's assessment of testing in biological practice reveals that static data are often insufficient to confirm hypotheses about process. Also, she makes the further claim that fit provides only "indirect" evidence for the assumptions of a model, and that independent "direct" tests of the assumptions are essential for confirmation in evolutionary biology (1988, 148). The independent assumptions of a model include specification of parameters, such as selection coefficients or population size, as well as other assumptions, such as random mating or Mendelian transmission. These parameters and assumptions represent aspects of the evolutionary dynamics. Thus Lloyd's take coheres with Lewontin's: static data often do not suffice to confirm hypotheses about the dynamic process. More "direct" evidence, the dynamic data, is needed. If Lloyd is right about testing then it is unsurprising that fit between a model and static data often faces a contrast failure problem with respect to hypotheses about evolutionary dynamics.

There are, embedded in both Lewontin and Lloyd, two ways to respond to the general threat of contrast failure with respect to static data. The first line of resistance denies that evolutionary biology has access to only static data, and insists that independent tests of the assumptions about dynamics are possible within our limitations. One strategy for providing dynamic data is to conduct long-term studies over many generations, as Lewontin admits above. The Grants' study provides one of the best examples of this, and given time and resources this sort of strategy can be applied to many different organisms. Based on the unpredictable fluctuations in selection differentials that Grant & Grant (2002) observed over 30 years, Lewontin adds a further worry that even these long-term studies fail to detect the nuanced evolutionary forces. So multiple generation studies may still have limits, but they approach the standard for dynamic data. Another more effective strategy, experimental manipulations in both laboratory and natural populations, provides support for the dynamical parameters incorporated into the model, such as selection coefficients, population sizes, and mutation rates. For selection models, manipulations can test whether certain ecological factors contribute to fit-

ness differences by investigating how populations respond to changes in those factors. This evidence, plus information about the historical presence of such ecological conditions, also approaches the standard for dynamic data, and the consensus is that such data provide the best evidence for selection in the wild (Endler, 1986). Lewontin should agree, as he suggests that the successful estimation of process parameters explains how studies of the rare model systems get at the dynamics (2002, 9). Although we may have only a few good cases so far, that we have any undermines general pessimism regarding our access to dynamic data.

The second line of resistance argues that many evolutionary studies that rely on static data, when carefully designed with precise hypotheses, can provide evidence that discriminates between evolutionary rivals and hence overcome contrast failure. Following Lloyd, we should recognize that fit generally provides *some* (“indirect”) evidence for dynamic assumptions rather than *no* evidence. The simple correlation study discussed above, when deployed in circumstances where we have rich background information about the target population, can provide evidence for selection over drift. In molecular evolution there are comparative tests that have provided evidence for positive selection on specific protein-coding genes. The MacDonald & Kreitman (1991) study is widely regarded as one of the best examples of molecular evidence for selection, though the MK test has definite limitations and can provide misleading results under some conditions (Eyre-Walker, 2002). The key to successfully using static data to confirm hypotheses about evolutionary dynamics is to make the alternatives precise by drawing upon rich information about the biological context, and to construct a test that will discriminate between them. While Lewontin is correct to emphasize the difficulties that face evolutionary inquiry, well constructed tests can turn static data into good evidence for rival hypotheses about dynamics.

In short, the *Spandrels* paper offers a specific critique of adaptationist testing methods. From the specific critique we can articulate a general warning for reconstructing evolutionary dynamics: the problem of contrast failure is a pervasive threat, and so confirmation in evolutionary biology requires more than just fitting static data. We can give normative force to Lloyd’s claim that evidence for the independent assumptions of evolutionary hypotheses is *necessary*; we need to meet the standards for dynamic data. Yet the warning is compatible with success in precise cases. Analyzing when testing methods overcome contrast failure, and

so enable inferences about dynamics, helps identify when biologists have strong evidence for their evolutionary hypotheses. Such evaluations will depend on the details of the specific testing problems biologists construct. The Lewontin and White case study will provide an example of how this evaluation should go. But first let me locate contrast failure in relation to another general problem of evidence.

### 3 Contrast failure and underdetermination

The problem of contrast failure is a distinctive type of *underdetermination*. Underdetermination encompasses a family of related problems that confront theory choice. Dietrich & Skipper (2007) argue that this family is large. They provide a comprehensive account of possible theory assessment strategies and generalize underdetermination to apply to choices made between alternatives on the basis of epistemic, pragmatic, and sociological desiderata. Since the focal issue in the Spandrels paper concerns the sufficiency of evidential support for historical hypotheses, I will discuss versions of underdetermination with the aim of isolating the problem that occurs when one epistemic desideratum—evidence—fails to guide a choice between alternative hypotheses.

One familiar version, call it classical underdetermination, concerns the empirical equivalence of theories by all in-principle evidence. This problem occurs when two alternative and incompatible theories make different commitments about the constitution of the world but share all the same empirical consequences. They are, in Earman's (1993, 21) words, "empirically indistinguishable" because *no* amount of evidence can discriminate between them. Some argue that classical underdetermination causes serious trouble for scientific realism (Van Fraassen, 1980; Earman, 1993; Turner, 2007), whereas others claim that this is merely a philosophical worry that our best scientific theories overcome (Laudan & Leplin, 1991; Kitcher, 1993).

Turner (2005) describes another version that he calls *local* underdetermination of hypotheses by *available* evidence. He argues that this sort of problem plagues historical sciences, such as evolutionary biology and geology, and defines it as follows.

A local underdetermination problem is any situation in which the following conditions are satisfied: (a.) Two incompatible

hypotheses, H and H\*, are genuine rivals. (b.) H and H\* are weakly empirically equivalent. (c.) As best anyone can tell, H and H\* have roughly equal portions of nonempirical theoretical virtue (simplicity, explanatory power, and the like). (d.) Background theories give us some reason to think that H and H\* are also strongly empirically equivalent (Turner, 2005, 218)

Turner's definition of local underdetermination turns on the distinction between weak and strong empirical equivalence. The distinction tracks whether certain total bodies of evidence fail to discriminate between the rivals. On Turner's account, H and H\* are weakly equivalent "if and only if they are both equally well supported by all the available evidence," whereas they are strongly equivalent "just in case they are (or would be) equally well supported by all the empirical evidence that will ever be available to us" (2005, 217). He then points to examples of local underdetermination in paleontology and geology (2005, 220–221).<sup>4</sup>

Turner's problem of local underdetermination bears a strong similarity to what Sklar (1975) calls *transient* underdetermination. This latter type of underdetermination occurs when there are "incompatible alternatives between which no rational choice can be made on the basis of a priori plausibilities, strength, simplicity, inductive confirmation, and so forth, *relative to present empirical evidence*" (Sklar, 1975, 380–381). The crucial difference between Sklar's transient version and Turner's local version involves a bet about future evidence. Transient underdetermination requires incompatibility between rivals, equal theoretical virtue, and empirical equivalence "relative to present empirical evidence." Local underdetermination requires all this (conditions (a) through (c)) plus good theoretical reasons for taking the empirical bet that data available to us in the future will continue to fail to discriminate rivals (condition (d)).

The trouble with the local version of underdetermination, in sharp contrast to the transient version, is that this empirical bet about the future set of evidence is simply too difficult to evaluate. We cannot be in an epistemic position *now* to assess whether two incompatible rivals are empirically equivalent relative to all *present and future* evidence. Turner argues that geology and paleontology give us two reasons to suspect condition (d) is met: natural processes (such as erosion or natural selection) tend to destroy information, and the historical record of prehistory tends to be incomplete (2005, 223–224). However, there are concerns that mitigate each.

First, as Sober (1988, 3–4) argues, historical processes *need not* be information destroying. Under some conditions natural selection can preserve information about ancestral states, and we simply cannot know how frequently these conditions are met before investigating specific evolutionary systems. Second, incompleteness of the geological or fossil record is an epistemic evaluation based on the *current state* of science. We can make no reliable inferences about how our epistemic position will persist into future. This leaves sufficient space to doubt that incompatible rivals will remain empirical equivalent after the accumulation of more data, the development of new technology, and the innovation of theory. But regardless of the plausibility of condition (d), both local and transient underdetermination clearly present more pressing problems for scientific inquiry than their classical relative (Stanford, 2006).

The problem of contrast failure belongs in this family of underdetermination problems but is a distinct relative. Classical underdetermination poses a problem when no in-principle evidence can distinguish between two theories. Contrast failure allows for empirical differences between two evolutionary rivals. It becomes a threat when some specific data set or contrastive test fails to detect these differences. Local and transient underdetermination are closer relatives to contrast failure since both focus on incompatible hypotheses that make different empirical commitments. Yet these types of underdetermination are assessed relative to *all* currently accessible data, and perhaps all data accessible in the future as well. Local or transient underdetermination occurs when that body of data fails to support one rival over another. Assessing contrast failure requires no bets about the evidential import of future data, nor does it depend on the complete set of currently available data. Instead, contrast failure identifies a more precise epistemic problem, one that frequently occurs and is often overcome in daily scientific practice. *A contrast failure problem occurs when a specific data set fails to discriminate between a defined set of evolutionary rivals.*<sup>5</sup>

To illustrate the problem of contrast failure and its differences from underdetermination consider the following idealized example from Sober (2005, 2008). Sober proposes a testing protocol to distinguish between directional selection (with some drift) and pure drift (no selection) for a simple quantitative phenotypic trait, the length of polar bear fur. The protocol postulates an optimality model that, based on a specific environment, identifies the optimum fur length for polar bears. Suppose we have such a model. Also suppose that fur measurements in the current population of

polar bears yield a mean that statistically coincides with the predicted optimum, and that a suitable amount of evolutionary time has elapsed such that if the ancestor of polar bears had a different fur length then selection will have had sufficient time to produce the optimum fur length. In this situation the observation that the present state of the population coincides with the optimum alone provides evidence for the selection (plus drift) hypothesis over the pure drift hypothesis (Sober, 2008, 200). This follows because that evolutionary outcome, where the present state of a population is statistically the same as the predicted optimum, is significantly more likely on the selection hypothesis. In this constrained case the static data, that the current population state statistically matches the optimum, successfully discriminates between selection (plus drift) and pure drift.

Let me alter the Sober's testing protocol by expanding the testing set. The *Spandrels* paper defends constraint as an alternative class of evolutionary hypotheses (Pigliucci & Kaplan, 2000). Consider a hypothesis that claims that the current state of the population is the product of a developmental constraint. Perhaps fur length is deeply entrenched in early development and the developmental program cannot change without severe maladaptive consequences. Suppose, for sake of argument, that the constraint hypothesis postulates an identical population distribution around the optimum value. Also suppose that the constraint hypothesis predicts that the ancestors of polar bears are capable of migration such that if some environmental change occurred, such as an increase in ambient temperature, then the bears would migrate to a colder habitat. The constraint hypothesis posits a different strategy, migration to a new environment rather than selection for changes in fur length, for a population of polar bears to reach the optimal trait-environment match. In this very idealized example the observation that the present state of the population statistically coincides with the putative optimum *fails* to discriminate between the selection (plus drift) and constraint (plus migration) hypotheses. Selection could have produced the optimum from a different ancestral state or constraint could have prevented the ancestral state from changing, prompting the bears to migrate to a habitat better suited to their constrained fur length. With respect to the new hypothesis set (selection plus drift and constraint plus migration) and the specific static data (the distribution of fur length in the current population) we have contrast failure. This particular testing protocol, testing whether the current population state matches the optimum, must thus be altered to provide evidence for selection versus

constraint.

Contrast failure reveals the risks these testing protocols face. The hypotheses, selection (plus drift), pure drift, and constraint (plus migration) are incompatible with obvious in-principle empirical differences. In realistic cases the empirical differences between evolutionary hypotheses about the length of polar bear fur will be numerous and at least some of these differences will be detectable in practice. In this idealized case, however, there is one key observation that will discriminate the selection and constraint hypotheses: the ancestral character state. If the relevant ancestral state is different from the current state then this counts as evidence for selection over constraint. If the ancestral state is identical to the current state then this confirms constraint over selection. Using phylogenetic methods and comparative data to determine the relevant ancestral trait values at the right places on the evolutionary tree overcomes contrast failure in this situation.

Identifying contrast failure problems helps diagnose the limitations of testing protocols. The other types of underdetermination lack the necessary precision. Transient and local underdetermination are assessed relative to a large body of evidence, whereas contrast failure is sensitive to the evidential import of specific data or observations. Evaluating underdetermination problems helps determine the overall support of particular theories, hypotheses, or sciences given the totality of (available) evidence and other evaluative criteria. Assessing the risk of contrast failure is a more useful task for guiding test construction in evolutionary biology, for this activity shows how to augment testing protocols. To clarify how identifying contrast failure helps evaluate testing in scientific practice, let me turn to an example from population genetics.

## 4 Adaptive landscapes

Lewontin & White (1960) collect frequency data on chromosome inversions for *Moraba scurra* (a species of grasshopper) and construct a classic example of an *adaptive landscape*. Lewontin and White use adaptive landscapes to represent the relation between a population's genetic composition and mean fitness. The "landscape" is a mathematical topography with peaks of high mean fitness and valleys of low mean fitness. Natural selection, if unconstrained and frequency independent, should push

the population to peaks in the landscape, where mean fitness is (locally) maximized.<sup>6</sup> The adaptive landscape purports to summarize all the interacting ecological factors responsible for the fitness differences among variants of the focal trait in *M. scurra*. The philosophically interesting feature of Lewontin and White's study is the set of assumptions they use to generate the adaptive landscape from frequency data. Looking closely at their methods will uncover limitations in the testing protocol. As will become clear, there are incompatible hypotheses that appeal to different combinations of demographic, genetic, and ecological factors to predict the same evolutionary outcome. Matching the predicted outcome with the observed state of the population does not discriminate between these rival hypotheses—the study faces several contrast failure problems. We can overcome these problems by obtaining evidence for the independent assumptions of the rivals.

Investigating any case of confirmation in evolutionary biology inevitably raises issues about idealization. All evolutionary models make some idealizations. The rival hypotheses in this case study are no exception. They involve highly idealized population genetic models that simply cannot represent any real evolutionary system in all respects. Yet this does not preclude the possibility of confirming these models as adequate or explanatory representations of some real system. Indeed, accruing evidential support is one way to demonstrate the legitimacy of specific idealizations. I will restrict my focus to the process of *model testing*, and how we should revise testing protocols when contrast failures occur. This process can be usefully disentangled from *model building*. Constructing a model involves assessing tradeoffs and making pragmatic decisions about what idealizations to include. There are rich accounts of how this construction process can and should go (Levins, 1966; Odenbaugh, 2005; Weisberg, 2007; Plutynski, 2007). Once we have the models built, we need to test them, and assess whether certain types of data discriminate between some set of rival models of a real system. While these activities certainly interact—idealizations will come under scrutiny as a result of model testing and inform future instances of model building—separating the genesis of idealized models from the confirmation of those models helps gain traction on the issue of evidential support.

Getting to the philosophical problems requires understanding the genetic details and the method for constructing the landscapes. Lewontin and White focus on two inversions on different chromosomes. *M. scurra*

are diploid (chromosomes come in pairs) with two possible alleles for each chromosome pair (an inversion type and a standard type), so there are nine possible genotypes. The data obtained are the frequencies of the nine genotypes in natural populations. The correlations between the actual frequencies of genotypes provide evidence that there is an epistatic interaction between the two chromosome pairs; the genotype of one chromosome pair affects the fitness of various types of the second pair. Given that there is an interaction, Lewontin and White want to explain why natural populations exhibit the particular gene frequencies they do (1960, 117). To do this they estimate fitnesses for the different genotypes and construct an adaptive landscape.

Estimating the fitnesses and constructing the landscape require making assumptions about the genetics, demography, and ecology. The fitness estimates are based on survival and are determined by comparing the actual adult genotype frequencies (post-selection) with postulated gamete or juvenile genotype frequencies (pre-selection). The postulated genotype frequencies depend on the assumptions of the Hardy-Weinberg equilibrium model (HWE). The assumptions include random mating, independent assortment of chromosomes, frequency independence, and that the ecological conditions have been stable long enough for a population to reach the gene frequency equilibrium due to selection (Lewontin & White, 1960, 118–119). The standard HWE assumptions allow Lewontin and White to calculate the postulated *genotype* frequencies from the overall *gene* frequencies observed in the populations. The viability fitness ( $W_i$ ) for each genotype  $i$  is given by the ratio of actual (observed) to postulated genotype frequencies (given in Table 1):  $W_i = p_{\text{actual}}/p_{\text{postulated}}$ . Lewontin and White can postulate the pre-selection *genotype* fitnesses in this way because of their assumption that the populations have reached equilibrium. At equilibrium the change in *gene* frequency due to selection equals zero. The HWE assumptions, including the assumption that gene frequencies are stable at equilibrium, entail that each generation should produce genotype frequencies that approximate HWE before viability selection acts. Also, the assumption of independent assortment entails that each genotype frequency at HWE equals the product of the HWE ratios for each locus. Deviations between the postulated HWE genotype frequencies and the actual, observed genotype frequencies are due, by hypothesis, to genotypic fitness differences that exist at equilibrium gene frequencies.

The adaptive landscape is generated from these viability fitness esti-

Table 1: CD and EF are the two chromosomes under study. Let  $p_a$  be the frequency of the standard CD chromosome ( $St$ ),  $q_a$  be the frequency of the CD inversion ( $Bl$ ),  $p_b$  be the frequency of the standard EF chromosome ( $St'$ ) and  $q_b$  be the frequency of the EF inversion ( $Tid$ ). (As is standard in population genetics,  $q_a = 1 - p_a$  and  $q_b = 1 - p_b$ .) The nine postulated genotype frequencies are obtained by multiplying the HWE ratios for each chromosome as shown here. At equilibrium  $\Delta p_a = \Delta p_b = 0$ .

	$St/St : p_a^2$	$St/Bl : 2p_aq_a$	$Bl/Bl : q_a^2$
$St'/St' : p_b^2$	$p_a^2p_b^2$	$2p_aq_ap_b^2$	$q_a^2p_b^2$
$St'/Tid : 2p_bq_b$	$2p_a^2p_bq_b$	$4p_aq_ap_bq_b$	$2q_a^2p_bq_b$
$Tid/Tid : q_b^2$	$p_a^2q_b^2$	$2p_aq_aq_b^2$	$q_a^2q_b^2$

mates. The landscape represents mean fitness ( $\bar{W}$ ) as a function of a possible set of the nine genotype frequencies ( $Z_i$ ) and relative viabilities ( $W_i$ ). The mean fitness is the average of the genotype fitnesses weighted by genotype frequency:  $\bar{W} = \sum_{i=1}^9 Z_i W_i$ . Each hypothetical combination of the nine genotypes determines a specific point on the adaptive landscape. If a population exists in a “valley” rather than on a “peak” in the landscape then directional selection should operate, increasing mean fitness, assuming fitnesses are constant and frequency independent. The adaptive landscapes depend exclusively on the estimated viability fitnesses. These fitness values, and the ecological factors responsible, are assumed to be stable. Lewontin & White (1960, 122–123) generate landscapes for various natural populations and offer the example in Figure 1 as indicative of the general trend they observe. They note that all natural populations rest on a “saddle point” in the adaptive landscape rather than on a peak. The data do not match the predictions of the model.

The mismatch between observation and prediction prompts a search for rival hypotheses capable of explaining the data. Recall that the adaptive landscape supposedly represents all the ecological interactions responsible for the fitness differences between genotypes. Also, these adaptive landscapes depend entirely on viability fitnesses estimated from frequency data. Given that the observed state of the populations does not fit the predictions of the model there must be some aspect of ecology (or de-

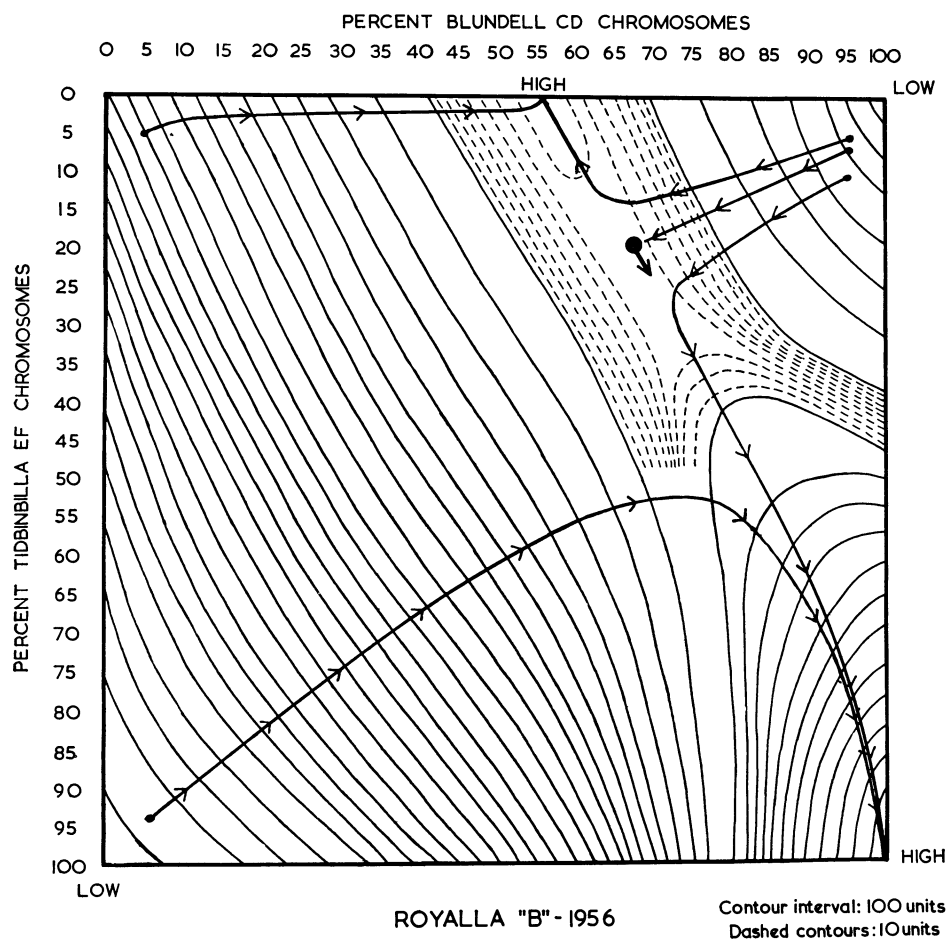


Figure 1: The adaptive landscape from Royalla B 1956 data used by Lewontin & White (1960, 127). The arrowed lines represent possible evolutionary trajectories across the landscape. The dot in the saddle point represents the observed state of the Royalla B population. The axes represent the frequencies of the inversions at each chromosome. "High" and "low" refer to peaks and valleys in the adaptive landscape respectively.

mography or genetics) that the model fails to take into account. Lewontin & White (1960, 125–126) suggest four alternative explanations for the mismatch: (1) the ecological conditions underlying the landscape have not existed long enough for the population to reach one of the peaks (the current state is what they call a “historical relic”); (2) the fitnesses used measure only viability and fail to include the fecundity component of fitness; (3) the fitnesses are frequency-dependent; and (4) yearly ecological variations change the landscape and keep natural populations at the saddle point.

Notice that each alternative specifies a confounding ecological condition. If (1) holds then the assumption made to generate the viability fitnesses is false. The population would not have reached equilibrium, and so the change in *gene* frequencies each generation due to selection will be positive or negative, not zero. Using HWE ratios to determine postulated genotype frequencies (as in Table 1) will not provide good estimates of the genotype viability fitnesses beyond the next generation. This entails that the adaptive landscape is only correct for the current generation; the population state as well as the landscape will change in subsequent generations. Alternatives (2) and (3) take the adaptive landscape to provide a good approximation of *viability* fitness but posit ecological complexities that would make the apparent saddle point an equilibrium state (a peak in the complete adaptive landscape). The last alternative (4) suspends the assumption of a constant environment and posits a different ecological pattern responsible for the fitness structure: wet and dry yearly oscillations change the local region of the landscape such that the population fails to escape the saddle point. Also notice that the available data fails to discriminate between any of the four ecological possibilities. There is contrast failure for the set of ecological rivals and the static data, although the study never aimed to overcome this particular problem.

While Lewontin and White identify rival hypotheses by considering possible ecological confounds, they do not consider suspending key assumptions about demography (random mating) or genetics (independent assortment) necessary to determine the postulated genotype frequencies. As Wright’s (1978, 127–145) review of this case makes clear, changing either of the demographic or genetic assumptions changes the postulated genotypes frequencies and thus the estimated viability fitnesses. Perhaps the data can discriminate between the original landscapes and rival landscapes generated from different assumptions.

Turner (1972) questions the assumption of independent assortment be-

cause epistatic interactions exist between the loci in *M. scurra*. This usually causes some degree of linkage disequilibrium. Linkage causes a deviation from the composite HWE ratios used to calculate the postulated genotype frequencies, and so an additional parameter, quantifying the degree of linkage between the two chromosomes, is necessary to calculate the viability fitnesses. Wright (1978), following Allard & Wehrhahn (1964), questions the assumption of random mating because *M. scurra* live in a fragmented habitat and have low dispersal ranges. These factors tend to produce some degree of inbreeding. Inbreeding also causes deviations in HWE ratios and requires an additional parameter to determine the postulated genotype frequencies. When the viability fitnesses and landscapes incorporate one of the additional parameters recommended by Turner and Wright the saddle points occupied by most populations turn into peaks (see Figure 2); observation and prediction now match.

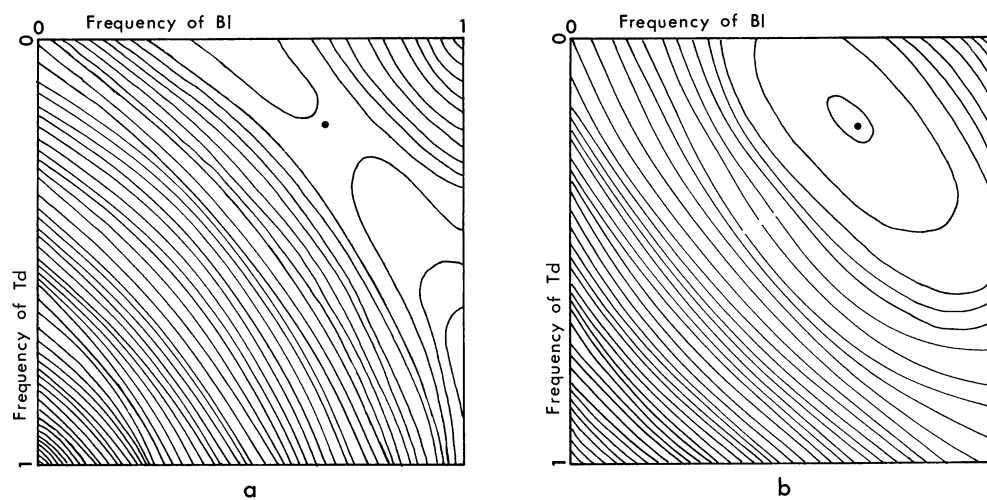


Figure 2: Two adaptive landscapes from the same Royalla B 1956 data (cf. Figure 1) adapted from Turner (1972, 336). Landscape (a) uses Lewontin and White's set of assumptions, and landscape (b) uses Turner's set of assumptions and a linkage disequilibrium parameter estimated from the same data. The black dot represents the observed state of the population. In (a) the population occupies a saddle point, whereas in (b) the population occupies a peak.

The evidential import of the data, however, is far from clear. The rival *M. scurra* adaptive landscapes face two contrast failure problems with respect to the observed genotypic state of the populations. First, contrast failure clearly occurs for the comparison between the Turner and Wright landscapes. Both alternatives locate peaks in the same place and predict that the population states will match. The data fits the predictions of both rivals, and so cannot discriminate between the genetic hypothesis (linkage) and the demographic hypothesis (inbreeding). Second, the risk of contrast failure undermines the comparison between the Turner-Wright landscapes and the Lewontin and White landscapes. *Prima facie*, the data do discriminate between these rivals—the observed population states do not match the predictions of the Lewontin and White landscapes—but the problem of contrast failure is not overcome for the following reasons.

Assessing whether the static data (observed genotype frequencies) favor one rival (Turner or Wright) over the other (Lewontin and White) is too sensitive to the particular fine-grained view of evidential favoring one adopts. The Turner and Wright alternatives do fit the observed evolutionary outcomes better; the populations reside on peaks, just as predicted, whereas they reside on the saddle points of the Lewontin and White landscapes. Yet the increased fit comes at a cost. Both Turner and Wright use additional parameters to determine the viability fitnesses. In general, increasing the number of parameters in a model increases the risk of *overfitting* (Forster & Sober, 1994; Burnham & Anderson, 2002; Hitchcock & Sober, 2004). If we judge the Turner-Wright versus Lewontin and White comparison based on simple fit then the data count as evidence for the former. But if we balance the increased complexity against fit, thus taking the risk of overfitting seriously, then that evidential support is probably artificial, or at least significantly reduced.

Another related factor contributes to the contrast failure problem for the Turner-Wright versus Lewontin and White comparison. Neither Turner nor Wright have independent evidence for their estimates of linkage disequilibrium and inbreeding. Instead, they determine some plausible value by appealing to general evolutionary trends or mining the same data set. Only a thin plausibility argument supports the proposal to increase the complexity of the model. Furthermore, they do not provide any evidence against the ecological confounds Lewontin and White propose. This lack of independent support contributes to contrast failure for this second comparison. The frequency data cannot adjudicate between the possible expla-

nations that appeal to Turner's genetic or Wright's demographic factors and those of comparable complexity that appeal to Lewontin and White's ecological factors. *What is needed is additional evidence for the genetic, demographic, and ecological assumptions made by the rival hypotheses.* Certainly this sort of additional evidence is hard to obtain, but it is necessary to overcome contrast failure. Further study over several generations, in the form genetic analyses to estimate the degree of linkage, demographic surveys to assess mating patterns and dispersal to estimate the degree of inbreeding, or ecological studies to determine the presence of the potential ecological confounds, would provide the necessary evidence.

Let me summarize the case study. Lewontin and White make idealizations about genetics, demography, and ecology to estimate viability fitnesses, generate an adaptive landscape, and plot the predicted evolutionary trajectories of populations over time. They found a mismatch between model and data and identified possible ecological causes for the mismatch. Alternatively, Turner and Wright changed the genetic and demographic assumptions to generate landscapes that match the data. In line with the methodological prescriptions in the *Spandrels* paper, they propose more complex evolutionary hypotheses that incorporate different non-adaptive assumptions. But the observed outcome data, genotype frequencies, simply fail to discriminate between the Turner and Wright hypotheses, and the risk of over-fitting plus the lack of independent evidential support undermines the contrast between the Turner-Wright and Lewontin and White hypotheses.

Yet the rival hypotheses have empirical differences that can be detected in practice. Independent evidence for or against the conjectured ecological confounds, the postulated degree of inbreeding, or the amount of linkage would help overcome the contrast failure problems. Such evidence is necessary to approach Lewontin's standard for dynamic data, and provide an explanation of the evolutionary dynamics of the two chromosomes. Accomplishing this task requires demonstrating that the idealized landscape adequately and appropriately represents the genotype fitnesses without merely incorporating extra parameters to accommodate the evolutionary outcome. The prescriptive recommendation follows from the need to overcome the problem of contrast failure.

## 5 Conclusion

Evolutionary inquiry must overcome problems of contrast failure, problems that occur when some specific set of data fails to discriminate between precise evolutionary rivals. Investigating how testing protocols in evolutionary biology are vulnerable to contrast failure provides one strategy for normative evaluation of methodology in evolutionary biology. This is part of the *Spandrels* attack on adaptationism. If the adaptationist program does not consider alternative evolutionary hypotheses, hypotheses capable of explaining the same outcomes, then it will fail to uncover adequate evidence for any hypothesis. Adequate evidence would be data that overcome contrast failure relative to both adaptation and non-adaptation rivals. The central importance of identifying when this pervasive problem of evidence occurs is the enduring methodological moral for testing evolutionary hypotheses that we should draw from the *Spandrels* paper.

The case study, Lewontin and White's adaptive landscapes of *M. scurra*, shows how to diagnose contrast failure problems and how to overcome them. Moreover, it shows that serious contrast failure problems emerge when we consider evolutionary rivals that concern possibilities other than direct selection. The diagnosis of contrast failure thus gives normative force to Lloyd's claim: confirmation in evolutionary biology requires more than just fit between predicted outcomes and static data. We must provide independent evidence for the many assumptions of our evolutionary hypotheses, evidence that approaches the standard for dynamic data. The case study also illustrates that this evidence is not impossible, or even difficult, to obtain. There are obvious ways to extend the study of *M. scurra* to make contact with the ecological, demographic, and genetic assumptions of the rival hypotheses. In the face of limitations on our epistemic access to evolutionary history, we should not surrender to pessimism. Instead, we should evaluate how to press evolutionary inquiry forward within such limits.

## Acknowledgements

Thanks to Peter Godfrey-Smith, Elliott Sober, Kyle Stanford, Kim Sterelny, Derek Turner, Ben Jeffares, and the audience at ISHPSSB 2005 for astute

comments and discussion. Thanks also to Bill Wimsatt for pointing me towards Wright's discussion of the *M. scurra* case.

## Notes

<sup>1</sup>This is but one of many threads woven into the *Spandrels* paper. A nexus of related adaptationist theses have been disentangled (Sober, 1996; Godfrey-Smith, 2001; Godfrey-Smith & Wilkins, 2008; Lewens, 2009), and there are proposals on how to test the empirical ones (Orzack & Sober, 1994, 1996; Brandon & Rausher, 1996). I will not focus on the standard issues of adaptationism, and instead turn to the general morals we can draw for testing any evolutionary hypothesis.

<sup>2</sup>That set of alternatives may include one or two key rivals (Royall's (1997) likelihoodism or Sober's (1990) contrastive empiricism) or the exhaustive list of possibilities (Bayesianism). Given that scientists seldom consider all possible hypotheses, instead focusing on a set of key rivals, I will construct my argument based on the more restrictive approach of Royall and Sober. My analysis can easily be embedded in a Bayesian framework by examining when two or more hypotheses (the genuine rivals) confer the same likelihood on some data.

<sup>3</sup>Beatty (1984, 196) notes the significance of this problem early on with regard to selection and drift: "... it is difficult to distinguish between random drift on the one hand, and the *improbable results of natural selection* on the other hand. Wherever there are fitness *distributions* associated with different types of organisms, there will be *ranges* of outcomes of natural selection."

<sup>4</sup>There is an important promissory note in Turner's analysis: he relies upon the concept of a genuine rival to define a local underdetermination problem, yet an account of genuine rivalry is not provided. Without such an account the problem of local underdetermination could be circumvented in any particular case by denying genuine rival status to an alternative, especially if scientists claim, reasonably in my view, that genuine rivals must admit of some in-practice discriminating evidence.

<sup>5</sup>Using the Dietrich & Skipper (2007, 303) framework, a contrast failure problem occurs when the X-set is the set of rivals for explaining a particular target evolutionary phenomenon, the Y-set is the set of data or observations collected from the target system, and the C relation that holds equally between the Y-set and each hypothesis in X is the confirmation relation. In other words, contrast failure occurs when a choice between precise evolutionary rivals (the X-set) is Dietrich-Skipper underdetermined by the evidence (the Y-set) with respect to purely epistemic evaluation (identical confirmation relations between the Y-set and each X).

<sup>6</sup>There are different ways to understand Wright's concept of a landscape. The landscape may map possible genotypes to genotypic fitness, or it may map possible population states to mean fitness. Lewontin and White take adaptive landscapes to do the latter. See Gavrillets (2004) for discussion of different evolutionary landscape concepts. Also see Wilkins & Godfrey-Smith (2009) for an insightful discussion of the utility of the landscape metaphor.

## References

- Allard, R. W. & Wehrhahn, C. (1964). A theory which predicts stable equilibrium for inversion polymorphisms in the grasshopper, *Moraba Scurra*. *Evolution*, 18, 129–130.
- Beatty, J. (1984). Chance and natural selection. *Philosophy of Science*, 51, 183–211.
- Beatty, J. (1987). Natural selection and the null hypothesis. In J. Dupre (Ed.), *The Latest on the Best* (pp. 53–75). MIT Press.
- Brandon, R. (1990). *Adaptation and Environment*. Princeton: Princeton University Press.
- Brandon, R. & Rausher, M. D. (1996). Testing adaptationism: A comment on Orzack and Sober. *American Naturalist*, 148, 189–201.
- Burnham, K. P. & Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach* (2nd ed.). New York: Springer.
- Dietrich, M. & Skipper, R. A. (2007). Manipulating underdetermination in scientific controversy: The case of the molecular clock. *Perspectives on Science*, 15(3), 295–326.
- Earman, J. (1993). Underdetermination, realism, and reason. *Midwest Studies in Philosophy*, 18, 19–38.
- Endler, J. (1986). *Natural Selection in the Wild*. Princeton: Princeton University Press.
- Eyre-Walker, A. (2002). Changing effective population size and the MacDonald-Kreitman test. *Genetics*, 162, 2017–2024.
- Fitelson, B. (2007). Likelihoodism, bayesianism, and relational confirmation. *Synthese*, 156, 473–489.
- Forster, M. R. & Sober, E. (1994). How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions. *British Journal of Philosophy of Science*, 45, 1–35.

- Gavrilets, S. (2004). *Fitness Landscapes and the Origin of Species*. Princeton: Princeton University Press.
- Godfrey-Smith, P. (2001). Three kinds of adaptationism. In S. H. Orzack & E. Sober (Eds.), *Adaptationism and Optimality* (pp. 335–357). Cambridge UP.
- Godfrey-Smith, P. & Wilkins, J. F. (2008). Adaptationism. In *The Blackwell Companion to the Philosophy of Biology*.
- Gould, S. J. & Lewontin, R. C. (1979). The spandrels of san marco and the panglossian paradigm: A critique of the adaptationist programme. *Proceedings of the Royal Society London, Series B, Biological Sciences*, 205, 581–598.
- Grant, P. R. (1986). *The Ecology and Evolution of Darwin's Finches*. Princeton: Princeton University Press.
- Grant, P. R. & Grant, B. R. (1989). *Evolutionary Dynamics of a Natural Population: The Large Cactus Finch of the Galapagos*. Chicago: University of Chicago Press.
- Grant, P. R. & Grant, B. R. (2002). Unpredictable evolution in a 30-year study of Darwin's finches. *Science*, 296, 707–711.
- Grant, P. R. & Grant, B. R. (2006). Evolution of character displacement in Darwin's finches. *Science*, 313, 224–226.
- Hitchcock, C. & Sober, E. (2004). Prediction versus accommodation and the risk of overfitting. *British Journal for the Philosophy of Science*, 55, 1–34.
- Kitcher, P. (1993). *The Advancement of Science*. New York: Oxford University Press.
- Laudan, L. & Leplin, J. (1991). Empirical equivalence and underdetermination. *Journal of Philosophy*, 88, 269–285.
- Levins, R. (1966). The strategy of model building in population biology. *American Scientist*, 54, 421–431.
- Lewens, T. (2009). Seven types of adaptationism. *Biology and Philosophy*.

- Lewontin, R. C. (2002). Directions in evolutionary biology. *Annual Review of Genetics*, 36, 1–18.
- Lewontin, R. C. & White, M. J. D. (1960). Interaction between inversion polymorphisms of two chromosome pairs in the grasshopper, *Moraba Scurra*. *Evolution*, 14, 116–129.
- Lloyd, E. A. (1988). *The Structure and Confirmation of Evolutionary Theory*. New York: Greenwood Press.
- MacDonald, J. & Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in drosophila. *Nature*, 351, 652–654.
- Odenbaugh, J. (2005). Idealized, inaccurate but successful: A pragmatic approach to evaluating models in theoretical ecology. *Biology and Philosophy*, 20, 231–255.
- Orzack, S. H. & Sober, E. (1994). Optimality models and the test of adaptationism. *American Naturalist*, 143, 361–380.
- Orzack, S. H. & Sober, E. (1996). How to formulate and test adaptationism. *The American Naturalist*, 148(1), 202–210.
- Pigliucci, M. & Kaplan, J. (2000). The fall and rise of Dr. Pangloss: Adaptationism and the Spandrels paper 20 years later. *Trends In Ecology and Evolution*, 15(2), 66–70.
- Plutynski, A. (2007). Strategies of model building in population genetics. *Philosophy of Science*, 73, 755–764.
- Royall, R. M. (1997). *Statistical Evidence: A likelihood paradigm*. New York: Chapman and Hall/CRC.
- Sklar, L. (1975). Methodological conservatism. *Philosophical Review*, 84, 384–400.
- Sober, E. (1988). *Reconstructing the Past: Parsimony, Evolution, and Inference*. MIT Press.
- Sober, E. (1990). Contrastive empiricism. In W. Savage (Ed.), *Scientific Theories*, volume 14 (pp. 392–412). Minneapolis: University of Minnesota Press.

- Sober, E. (1996). Evolution and optimality: Feathers, bowling balls, and the thesis of adaptationism. *Philosophic Exchange*, 26, 41–55.
- Sober, E. (2005). Is drift a serious alternative to natural selection as an explanation of complex adaptive traits? In A. O’Hear (Ed.), *Philosophy, Biology and Life*. Cambridge: Cambridge University Press.
- Sober, E. (2008). *Evidence and Evolution: The Logic Behind the Science*. Cambridge: Cambridge University Press.
- Stanford, P. K. (2006). *Exceeding Our Grasp: Science, History, and the Problem of Unconceived Alternatives*. Oxford: Oxford University Press.
- Turner, D. (2005). Local underdetermination in historical science. *Philosophy of Science*, 72, 209–230.
- Turner, D. (2007). *Making Prehistory: Historical Science and the Scientific Realism Debate*. Cambridge: Cambridge University Press.
- Turner, J. R. G. (1972). Selection and stability in the complex polymorphism of *Moraba Scurra*. *Evolution*, 26, 334–343.
- Van Fraassen, B. C. (1980). *The Scientific Image*. Oxford: Clarendon.
- Weisberg, M. (2007). Who is a modeler? *British Journal for the Philosophy of Science*, 58, 207–233.
- Wilkins, J. F. & Godfrey-Smith, P. (2009). Adaptationism and the adaptive landscape. *Biology and Philosophy*.
- Wright, S. (1978). *Evolution and the Genetics of Populations: Variability within and among Natural Populations*, volume 4. Chicago: University of Chicago Press.