WILEY-BLACKWELL

Cladistics 24 (2008) 1–23

Cladistics

10.1111/j.1096-0031.2008.00214.x

Parsimony and explanatory power

James S. Farris*

Molekylärsystematiska laboratoriet, Naturhistoriska riksmuseet, Box 50007 SE-104 05 Stockholm, Sweden

Department of Invertebrate Zoology, American Museum of Natural History, Central Park West at 79th Street, New York, NY 10024, USA

Accepted 24 December 2007

Abstract

Parsimony can be related to explanatory power, either by noting that each additional requirement for a separate origin of a feature reduces the number of observed similarities that can be explained as inheritance from a common ancestor; or else by applying Popper's formula for explanatory power together with the fact that parsimony yields maximum likelihood trees under No Common Mechanism (NCM). Despite deceptive claims made by some likelihoodists, most maximum likelihood methods cannot be justified in this way because they rely on unrealistic background assumptions. These facts have been disputed on the various grounds that *ad hoc* hypotheses of homoplasy are explanatory, that they are not explanatory, that character states are ontological individuals, that character data do not comprise evidence, that unrealistic theories can be used as background knowledge, that NCM is unrealistic, and that likelihoods cannot be used to evaluate explanatory power. None of these objections is even remotely well founded, and indeed most of them do not even seem to have been meant seriously, having instead been put forward merely to obstruct the development of phylogenetic methods.

Introduction

By the early 1980s parsimony was already well established as the method of choice among phylogenetic systematists, but the justification of the method still seemed incomplete. Kluge and Farris (1969), for example, had pointed out that a most parsimonious tree was the best fit to available characters, but this left open the question of why that particular measure of fit should be used. In 1983 I was able to resolve this issue, as well as several others, by relating parsimony to explanatory power, that is, by showing that parsimony assesses the degree to which a tree can account for observed similarities among terminals as the result of inheritance from a common ancestor (Farris, 1983). Later (Farris, 2000; Farris et al., 2001) I refined that idea by combining Popper's (1959) formula for explanatory power with a relationship between parsimony and likelihood that had been discovered by Tuffley and Steel (1997). Of course no good deed goes unpunished. My derivations have been criticized by opponents of parsimony (de Queiroz and Poe, 2003; Felsenstein, 2004), by opponents of Popper (Rieppel, 2003), and by inventive authors with their own distinctive theories on Popper (Faith, 1992, 2004) and phylogenetic evidence (Kluge and Grant, 2006). It can be beneficial to examine such objections, for this can help to clarify points that might otherwise have been incompletely understood. My purpose here, accordingly, is to explain why those criticisms are not well founded.

Explanatory power

My 1983 conclusion can be derived from a few simple ideas, the first of which concerns what genealogies can explain (Farris, 1983, p. 18):

> Genealogies provide only a single kind of explanation. A genealogy does not explain by itself why one group acquires a

*Corresponding author:
E-mail addresses: steve.farris@nrm.se

> new feature while its sister group retains the ancestral trait … A genealogy is able to explain observed points of similarity among organisms just when it can account for them as identical by virtue of inheritance from a common ancestor.

Of course such explanations would not apply to purely phenotypic variation. If (as I will assume throughout) phenotypic variation has been removed and any errors of observation have been corrected, similarities not explained by inheritance from a common ancestor are considered homoplasies, that is, cases of multiple origins of a feature. Some conclusions (hypotheses) of homoplasy can be supported directly by further investigation, for example by discovering that structures previously coded as alike are actually quite different, or even by corroborating a theory that would explain the particular multiple origins in question. Important though such possibilities may be in other contexts, they are of no interest here, for they mean that there is no inherited similarity for the tree to explain, and for purposes of this discussion I will assume that all such cases have already been eliminated. Homoplasies still remaining are those that are concluded simply because they are implied by the tree, so that they have no supporting evidence of their own. Such hypotheses are called *ad hoc* (Farris, 1983, p. 10):

> If a conflicting character survives all attempts to remove it by searching for such evidence, then the conclusion of homoplasy in that character, required by selecting a placement [of a terminal on the tree], satisfies the usual definition of an *ad hoc* hypothesis. It is required to defend the genealogical hypothesis chosen, but it is not supported by any evidence separate from that for the genealogy itself. If external evidence favors the interpretation of homoplasy, however, that hypothesis is not *ad hoc*.

Ad hoc hypotheses of homoplasy, then, correspond to observed similarities that are explained neither by inheritance from a common ancestor, nor, so far as is known, by anything else. They could simply be called unexplained similarities, and indeed this would often be clearer, although understanding "*ad hoc* hypotheses of homoplasy" is still necessary when discussing earlier literature.

As with any scientific theory, it is desirable for a tree hypothesis to explain as much of available observation as possible, and this means choosing the tree to minimize the number of similarities left unexplained. But some caution is needed when counting *ad hoc* hypotheses of homoplasy, for only the number actually required (implied) by the tree should be counted. Otherwise one could simply postulate superfluous homoplasies as "grounds" for criticizing any tree. This means that the homoplasies counted in evaluating a tree should be mutually independent, as otherwise some requirement might be counted more than once. It is common for homoplasies to be logically interdependent (Farris, 1983, p. 20):

> Suppose that a putative genealogy distributes [the 20 terminals showing feature X] into two distantly related groups A and B of ten terminals each. There are 100 distinct two-taxon comparisons of members of A with members of B, and each of those similarities in X considered in isolation comprises a homoplasy … [But if] X is identical by descent in any two members of A, and also in any two members of B, then the A-B similarities are all homoplasies if any one of them is.

But fortunately it is easy to count mutually independent homoplasies (Farris, 1983, p. 20):

> If a genealogy is consistent with a single origin of a feature, then it can explain all similarities in that feature as identical by descent. A point of similarity in a feature is then required to be a homoplasy only when the feature is required to originate more than once on the genealogy. A hypothesis of homoplasy logically independent of others is thus required precisely when a genealogy requires an additional origin of a feature. The number of logically independent *ad hoc* hypotheses of homoplasy in a feature required by a genealogy is then just one less than the number of times the feature is required to originate independently.

De Laet (2005) has arrived at the same result by another argument. To minimize independent unexplained similarities, one need only minimize required extra steps.

The "required" steps (or homoplasies) in that prescription are simply those that appear when characters are fitted to the tree, as in optimization (Farris, 1970). That a tree "requires" a certain homoplasy has sometimes been taken to mean that the similarity in question would falsify the tree, or at least that it would falsify the tree if the tree were not rescued *ad hoc* by the hypothesis of homoplasy. In fact no such interpretation is necessary for purposes of evaluating explanatory power. Unexplained similarities are simply that, and would not falsify the tree except perhaps on the bizarre assumption that homoplasy is impossible.

The later refinement (Farris, 2000; Farris et al., 2001) is based on Popper's (1959, p. 401) formula for the explanatory power E of hypothesis h with respect to evidence e, given background knowledge b, that is, the power of h to explain e (given b): [1]

$$E(h,e,b) = \frac{p(e,hb) - p(e,b)}{p(e,hb) + p(e,b)}$$

Most of the same comments will apply to Popper's (1983, p. 240; cf. Popper, 1963, p. 288; Popper, 1959, p. 400f) corroboration C of h by e (given b), which differs just in having an additional term in the denominator:

$$C(h,e,b) = \frac{p(e,hb) - p(e,b)}{p(e,hb) - p(eh,b) + p(e,b)}$$

Here $p(e, hb)$ denotes the probability of e given both b and h, while $p(e, b)$ is the probability of e given only b, that

[1] I have changed Popper's (1959, p. 401) symbols to h, e, b and p, respectively, to ease comparison with other formulae.

is, without h. In phylogenetic applications h would be a tree hypothesis and e would comprise the matrix of observed features of the terminals. As for background knowledge, Popper (1983, p. 236) explained:

> By background knowledge we mean any knowledge (relevant to the situation) which we accept—perhaps only tentatively—while we are testing h.

Similarly (Popper, 1963, p. 288):

> Here [b] should be taken as the general 'background knowledge' (the old evidence, and old and new initial conditions), including, if we wish, accepted theories.

Thus an evolutionary theory (or model) relied on when inferring phylogeny would be included in the background, though this would only be justified for accepted theories, by which Popper (1983, p. 243) meant:[2]

> If we mean by the degree of acceptability of a theory h the degree to which h is satisfactory from the point of view of *empirical science*—that is, from the point of view of *the aims of empirical science*—then acceptability will have to become topologically equivalent to corroboration.

Further (Popper, 1983, p. 230):

> 'When do we—tentatively—accept a theory?' Our answer is, of course: 'When it has stood up to criticism, including the most severe tests we can design; and more especially when it has done this better than any competing theory.'

This means that accepted theories must, as far as is known, be realistic. A theory that has been rejected has certainly not stood up to the most severe tests and can hardly be considered knowledge, background or otherwise.

As Popper (1959, p. 410) noted, the first term $p(e, hb)$ in these formulae is the likelihood of h given e (and b), and it is easily seen from the formula that $E(h, e, b)$ increases with $p(e, hb)$. As the second term $p(e, b)$ does not depend on h, the tree with greatest likelihood also has greatest $E(h, e, b)$. Purely algebraically, then, any maximum likelihood method would maximize the power to explain observations e if its model were included in b. That would be an abuse of Popper's formula, however, unless the model were accepted, and realism is seldom the primary aim of likelihoodists. Felsenstein (1993, in dnaml.doc), for example, confessed:

> This rather disconcerting model is used because it has nice mathematical properties which make likelihood calculations far easier.

The lack of realism in most present likelihood models (see Farris, 1999) arises from restrictive homogeneity assumptions. The ratio in expected substitution rates between any two characters is required to be the same in all branches of the tree. Sometimes, as in clock models, the rates are even required to be uniform. But Tuffley and Steel (1997) introduced a model called No Common Mechanism (NCM), in which characters may—but are not required to—vary their relative rates independently, both within and between branches. Because the independent variation is taken only as a possibility, not as a requirement, NCM would apply to almost any situation, and so may be accepted as realistic. This is useful because Tuffley and Steel also showed that maximum likelihood under NCM selects the same trees as does parsimony. With the realistic NCM in the background, then, most parsimonious trees have greatest power to explain available observations.

History

According to Felsenstein (2004), my earlier paper (Farris, 1983) had left important issues unsettled. Remarkably, he was able to draw that conclusion from a passage on p 8, the second page of my 30 pp paper (Felsenstein, 2004, p. 140):

> It [the hypothetico-deductive approach] is also invoked by Farris (1983, p. 8):
>
> > Wiley [(1975)] discusses parsimony in a Popperian context, characterizing most parsimonious genealogies as those that are least falsified on available evidence. In his treatment, contradictory character distributions provide putative falsifiers of genealogies. As I shall discuss below, any such falsifier engenders a requirement for an *ad hoc* hypothesis of homoplasy to defend the genealogy. Wiley's concept is then equivalent to mine.
>
> One might note that these discussions [the other was from Wiley (1981, p. 111)] do not distinguish clearly between parsimony and compatibility methods … When a character state arises three times on a phylogeny, the issue is whether we are to count that as one invalid character test [*ad hoc* hypothesis of homoplasy] or two, and whether the decision is implicit in the works of Popper, William of Ockham, or Hennig. This question is not directly dealt with in any of the philosophical writings of phylogenetic systematists. Phylogenetic systematists have tended to back parsimony and denounce compatibility. This seems to come, not from any philosophical principle, but from the feeling that compatibility discounts a character's value too rapidly, that there is still good information to be had from characters that have been observed to change more than once on a tree.

Of course this was a ruse. I did not address those issues on p 8, but later I showed exactly how to count *ad hoc* hypotheses of homoplasy, by counting extra steps. That derivation, quoted above, was on p 20. I also discussed clique methods, in the section titled "Cliques" on pp 30–33. There I demonstrated that the clique approach throws away explanation of observed similarities and in the process requires a preposterous assumption (Farris, 1983, p. 31; underlining added):

[2] As in quotations throughout, italics are as in the original.

> The covering assumption utilized is then that excluding a character [from the clique]—concluding that it shows somehomoplasy—implies that all points of similarity in that character are homoplasies.

That is readily recognized as the objection to cliques that—according to Felsenstein—came only from a "feeling". In that section I also pointed out (Farris, 1983, p. 32):

> Both [of Felsenstein's (1979, 1981)] rationalizations of the clique assumption, it will be seen, have the same defect as the clique assumption itself.

Perhaps this explains why Felsenstein wanted his readers to believe that I had not discussed cliques.

Evidently unwilling to address my comments, Felsenstein (2004) invented an alternate history in which I had never made them, the one quoted passage serving to mislead his readers further by creating the impression that he was discussing my paper forthrightly. As he also worked uncertainty about Hennig into his account, it seems pertinent to note that he had taken a similar approach before (Felsenstein, 1983, p. 323):

> I find it impossible to tell from a reading of Hennig whether he would have preferred parsimony to compatibility.

Parsimony can apply non-unique derivations as synapomorphies, but clique methods do not. It would then seem that Felsenstein could not tell from his reading whether Hennig would have applied non-unique derivations. But Hennig (1981, p. 23f) would have, and did:

> In such a [phylogenetic] classification, the fact that the Megasecoptera have acquired their neoptery independently means that they should not be placed in the Neoptera. This is because there are good reasons for believing that the neoptery of the Areoptera has arisen monophyletically and is one of the constitutive characters [synapomorphies] of this group. It is quite another matter that it has arisen elsewhere as well.

In fact he did so repeatedly, as Farris and Kluge (1986, p. 299) have summarized:

> Hennig (1983, p. 145) lists thecodont dentition as a synapomorphy of Archosauromorpha, even though he notes that it is convergently developed in Mammalia, and even though teeth are lost in some Archosauromorpha. On page 146 he lists bipedalism as a synapomorphy of the same group, although that trait is also developed in some Mammalia and secondarily lost in many Archosauromorpha. On page 135 he lists loss of teeth as a synapomorphy of Testudines, although that loss occurs also in Aves, as well as in other groups. Hennig, then, certainly does not restrict synapomorphy to similarity in traits that are uniquely derived.

Most readers would have taken "a reading of Hennig" to mean that Felsenstein had made a sincere effort to discover Hennig's views, and no doubt Felsenstein expected that interpretation. In fact by "a reading of Hennig" Felsenstein apparently meant the opposite.

Felsenstein (2004) seemed more direct in discussing my later (Farris, 2000) derivation. He reinvoked his earlier (Felsenstein, 1978) argument that parsimony can be statistically inconsistent, that is, can yield an incorrect tree when the data comprise infinitely many characters, randomly selected from the statistical population defined by a suitably chosen model. In his conception (Felsenstein, 2004, p. 143):

> Farris (1999, 2000) … argues that when a sufficiently realistic model [NCM] of variation of evolutionary rates among sites is adopted, parsimony obtains the same tree as likelihood and hence the tree favored by Popper's measure. I have already noted that in such cases the inference can be inconsistent. In such a case Popper's formula is corroborating the wrong tree!

But at this point he became less direct. In that case likelihood would also be selecting the wrong tree! Yet Felsenstein did not put it that way. Doing so would have conflicted with an earlier claim of his (Felsenstein, 1978, p. 408):

> Methods of phylogenetic inference which entirely avoid the problem of statistical inconsistency are already known. Maximum likelihood is one of them. I have outlined elsewhere (Felsenstein, 1973) how this may be done.

That claim was not true (see Farris, 1999) and Felsenstein (2004) was more circumspect, but he still wanted to associate likelihood with statistical consistency. He went on (Felsenstein, 2004, p. 143):

> If more is known about the distribution of evolutionary rates, one might be able to use a more specific model that achieved consistency.

But to be known—as distinguished from imagined—the distribution of rates would have to be established as realistic, and this raises another problem. Consistency can be assured only under relatively simple models (Steel et al., 1994; cf. Farris, 1999), so that a model complex enough to be realistic may be too complex for consistency. As I put it before (Farris, 1999, p. 203):

> Guarantees of consistency were never more than a sham issue. They can be achieved only under absurdly oversimplified circumstances. In the real world, no method can guarantee consistency. Nor should this be surprising. The idea that consistency can be guaranteed in real—as opposed to imaginary—cases is readily recognized as a version of a now-classical philosophical blunder, the belief that empirical inductions can be made infallible (for a discussion see Popper, 1972).

Felsenstein (2004) discussed this subject even less than he did my comments on cliques.

ASP

Kluge and Grant (2006) noted that minimizing steps no longer corresponds to minimizing homoplasies if the alignment is allowed to vary (as in POY; see Wheeler

et al., 2006). While that is well known, they took it to mean that an entirely new philosophical basis for parsimony would be needed and proceeded to offer their own. Their approach was based on the anti-superfluity principle (ASP), by which they meant that transformation events should be minimized (Kluge and Grant, 2006, p. 285):[3]

> The **h** that which [*sic*] minimizes the number of transformation events hypothesized to explain the character-states of terminal taxa as homologues is optimal.

This seemed much like my conclusion, but whereas my derivation concerned explaining observed similarities, they aimed to explain character states, and they had their own conception of character states (Grant and Kluge, 2004, p. 26):

> Character-states are defined conceptually as the least inclusive historical individuals that result from heritable transformation events.

If character states were historical individuals, no character state could originate more than once. What others might call two origins of the same state would have been, on Kluge and Grant's view, origins of different states individuated by historically distinct transformation events. That led to a defect in Kluge and Grant's (2006, p. 284) formulation of explanation:

> The *explananda*, **e**, are the character-states (sensu Grant and Kluge, 2004, [p. 26; quoted above]) of terminal taxa, which are explained by postulating a particular hypothesis of phylogenetic relationships (i.e. a hypothesis of cladistic and patristic relationships; Farris, 1967), **h**, in light of the background knowledge of descent, with modification, **b**. Together, **b** and **h** constitute the *explanans*.

The patristic part of hypothesis **h** consisted of transformations, and those were the same transformations that individuated the character states in **e**, so that **h** and **e** overlapped. Popper (1963, p. 288f; quoted in full later) had a rule against such practices:

> This rule for the exclusion of *ad hoc* hypotheses may take the following form: the hypothesis *must not repeat* (except in a completely generalized form) the evidence, or any conjunctive component of it.

Kluge and Grant's concept of character states thus seems to have led to violating Popper's rule for exclusion of *ad hoc* hypotheses, and they should have been aware of that rule, as Farris, Kluge and Carpenter (2001, p. 440) had called attention to it. But then why did Kluge and Grant adopt their concept of character states? Because (Grant and Kluge, 2004, p. 25):

> Character-states have generally been conceptualized as properties (attributes, features), which logically denies their ability to transform or evolve, since properties are class concepts and, as such, are immutable (Kluge, [2003]). Only individuals (in the ontological sense) can undergo change.

This was merely a confusion. Evolutionary transformation in a character simply means that an evolving population substitutes one state for another. A population may substitute a G for a T at some site, and this is in no way hindered by the fact that G and T themselves are distinguished by fixed chemical properties. Kluge and Grant's concept of character states, then, never had any legitimate motivation.

Unfortunately, their character state concept was not the only ill-motivated part of Kluge and Grant's discussion. They also commented on earlier treatments (Kluge and Grant, 2006, p. 278):

> Unlike Hennig (1966) and Farris (1980, 1983), Sober (1988, p. 33; see also Farris, 1967; Sober, 1986) took an explicitly probabilistic position on the presence of apomorphous similarities in different species, viewing
>
> > synapomorphies as providing evidence for monophyletic groups, rather than as absolutely guaranteeing that they must exist. A 110 character [A and B have apomorphic state 1, while C has plesiomorphic state 0.] does not deductively imply that A and B form a monophyletic group apart from C.

The suggestion that Hennig assumed such deductive implication—which would mean assuming impossibility of homoplasy—did not come from Sober, who knew better (Sober, 1988, p. 119):

> Hennig is obviously alive to the fact that shared derived characters may or may not reflect phylogeny.

Kluge and Grant should have known better too, since Farris and Kluge (1986, p. 299; quoted above) had already called attention to Hennig's use of multiply derived traits. It would seem that Kluge and Grant's suggestion said less about Hennig's ideas than about their own enthusiasm for attributing faults to others.

Kluge and Grant's comment about deductive implication included me as well, but in my case they were not content with suggesting that I thought homoplasy impossible. They also complained that I did hypothesize homoplasy (Kluge and Grant, 2006, p. 280):

> The problem we identify in Farris's (1983) rationale is not its focus on *ad hoc*-ness, but rather the premise that a statement of homoplasy entails an *ad hoc* hypothesis. Homoplasy cannot be an *ad hoc* hypothesis because it is not a hypothesis; it is an acausal description, not a causal explanation (Kluge, 1999).

There is actually no requirement that a hypothesis must be causal, but that question of usage hardly matters. Their comment could have been relevant as a substantive criticism only if I had claimed that *ad hoc* hypotheses of homoplasy were causal explanations, whereas in fact I associated those hypotheses with *un*explained similarities. That of course is why such

[3] Kluge and Grant (2006) used boldface rather than italics for Popper's symbols *h*, *e* and *b*.

hypotheses should be minimized, but Kluge and Grant (2006, p. 281) went on to find a problem with that minimization:

> Sober (1988, pp. 135–141) reiterated the long-standing criticism that minimizing an [*sic*] *ad hoc* hypothesis of homoplasy assumes that homoplasies are rare.

As this was their only mention of the minimization issue, they created the impression that minimization would pose a difficulty only in connection with *ad hoc* hypotheses of homoplasy, so leaving ASP minimization unthreatened. But Felsenstein (1984, p. 183) had phrased his accusation more broadly:[4]

> The general pattern is quite simple: if a method involves trying to find the tree that minimizes the number of occurrences of some evolutionary event, it implicitly assumes that the event is *a priori* improbable, so that its occurrence strains our credulity.

If that had been correct, it would have applied to any minimization of events, including that used in ASP, so that at best Kluge and Grant's account amounted to expressing an unfounded preference for their own approach. They also objected to explaining similarities (Kluge and Grant, 2006, p. 278):

> Darwin's theory of descent or common ancestry does not require an explanation of similarities among organisms.

It is already such an explanation, but they meant the opposite, going on (Kluge and Grant, 2006, p. 279):

> Similarities, by definition, cannot evolve. Similarity is specified in relation to one or more properties, which entails intentional definition. And, that being the case, the properties so defined must be considered immutable.

In my 1983 paper, similarities were treated as observations. Observations are not supposed to evolve, even when they reflect some recent state of populations that are themselves evolving. But this is only another version of the confusion underlying Kluge and Grant's character state concept. Remarkable though it seems, they apparently could not understand that populations can evolve simply by substituting one state for another.

Likelihood

Likelihoodists de Queiroz and Poe (2003) wanted to justify existing maximum likelihood methods as Popperian. Thus, although they never cited it, they agreed with my (Farris, 2000) observation that maximizing likelihood $p(e, hb)$ would maximize E and C. But as they meant to apply that result to existing maximum likelihood methods, they did not want to exclude unrealistic models from background knowledge. Their (de Queiroz and Poe, 2003, p. 356f) defense of that position was based on two passages from Popper (1963, p. 238):

> While discussing a problem we always accept (if only temporarily) all kinds of things as *unproblematic*: they constitute for the time being, and for the discussion of this particular problem, what I call our *background knowledge*. Few parts of this background knowledge will appear to us in all contexts as absolutely unproblematic, and any particular part of it *may* be challenged at any time, especially if we suspect that its uncritical acceptance may be responsible for some of our difficulties.

And

> The fact that, as a rule, we are at any given moment taking a vast amount of traditional knowledge for granted … creates no difficulty for the falsificationist or fallibilist. For he does not *accept* this background knowledge; neither as established nor as fairly certain, nor yet as probable. He knows that even its tentative acceptance is risky, and stresses that every bit of it is open to criticism, even though only in a piecemeal way.

Their interpretation of those passages emphasized that choice of background theories may depend on the problem being investigated, that acceptance is tentative, and that some background knowledge may have been accepted uncritically (de Queiroz and Poe, 2003, p. 356):

> This [first] statement, with its use of the phrases "if only temporarily," "for the time being," and "for the discussion of this particular problem," emphasizes the tentative nature of many aspects of *b*... Moreover, Popper's statement explicitly acknowledges that *b* may include components that have been accepted uncritically, which can hardly be equated with well-corroborated theories.

And (de Queiroz and Poe, 2003, p. 357):

> This [second] statement explicitly rejects the equation of *b* with accepted or established (well-corroborated) theories. Furthermore, it emphasizes that any tentative acceptance (i.e. for the purpose of a particular test) is risky, a caution that reinforces his earlier point that certain components of *b* may have been uncritically accepted and therefore are anything but well corroborated.

Of course it is true that choice of background depends on the problem, but that is a matter of relevance and hardly implies lack of realism, so that de Queiroz and Poe's emphasis on that dependence would seem to have been merely an attempt to create a misleading impression. Similarly, they wrote as if recognizing acceptance as tentative would separate acceptance from corroboration, but that was also misleading, for in fact Popper always regarded corroboration as a guide to tentative acceptance (Popper, 1983, p. 230):

> When do we—tentatively—accept a theory? Our answer is, of course: When it has stood up to criticism, including the most severe tests we can design; and more especially when it has done this better than any competing theory.

[4] If such reasoning were correct, I noted (Farris, 1983, p. 13), least-squares regression would have to assume that the residual variance is truly small. Interestingly, Felsenstein (2004) never mentioned his 1984 paper.

Further (Popper, 1959, p. 415; underlining added):

> As to degree of corroboration, it is nothing but a measure of the degree to which a hypothesis *h* has been tested… it is a measure of the rationality of accepting, tentatively, a problematic guess, knowing that it is a guess—but one that has undergone searching examination.

This leaves only their tacit suggestion—paradoxical, in view of their stress on tentative acceptance—that uncorroborated theories, once uncritically accepted, would somehow remain as background knowledge. Presumably de Queiroz and Poe meant to include the models used in likelihood methods as "traditional" knowledge. Their suggestion, however, was based on nothing but ignoring Popper's comment that any part of the background "may be challenged at any time, especially if we suspect that its uncritical acceptance may be responsible for some of our difficulties." Once challenged—critically tested—either a theory will be rejected, and so no longer accepted, or else it will be corroborated, in which case its tentative acceptance will no longer be uncritical. Any sincere attempt to eliminate difficulties will thus lead to selecting realistic—corroborated—background theories. de Queiroz and Poe's (2003) argument, then, amounted simply to the idea that scientific investigation should not be sincere.

Indeed, that idea seems to have played a significant role in the formation of de Queiroz and Poe's (2003) position, as can be seen from an earlier comment of theirs (de Queiroz and Poe, 2001, p. 315):

> According to Popper (1963, p. 238), "Few parts of the background knowledge will appear to us in all contexts absolutely unproblematic, and any particular part of it may be challenged at any time, especially if we suspect that its uncritical acceptance may be responsible for some of our difficulties" (see also Popper, 1983, p. 188). The provisional nature of background knowledge described by Popper allows phylogeneticists to evaluate not only alternative topologies but also alternative phylogenetic methods or models in terms of degree of corroboration.

In 2001 they understood correctly[5] that background theories should be corroborated; it was only later that they decided to ignore this part of Popper's comments. That shift was correlated with their treatment of my (Farris, 2000; Farris et al., 2001) discussion, which de Queiroz and Poe (2001) never mentioned, and which called attention to the consequences of Popper's acceptance requirement. When de Queiroz and Poe (2003) at last mentioned Farris et al. (2001), they had to try to hide the acceptance requirement in order to maintain their commitment to existing likelihood methods. They hoped to create the impression that their preference in methods was consistent with Popper's ideas, but of course what they actually showed was just the opposite.

That commitment to existing likelihood methods also seems to have underlain another part of de Queiroz and Poe's (2003) discussion. While otherwise they contended that unrealistic models could be used as background knowledge, there was one model to whose realism they did object, and that was the one related to parsimony, NCM (de Queiroz and Poe, 2003, p. 358):

> At least for some kinds of data, the assumption of no common mechanism seems more likely to be false. For example, for pseudogenes and most third-codon positions in protein-coding genes, where natural selection presumably does not affect the rate of substitution, it seems reasonable to expect that transformations in all characters have higher probabilities on branches of long temporal duration than on those of short temporal duration rather than the probabilities of change being entirely independent of the temporal duration of branches.

It is not just their double standard and their attempt to use their expectations in place of evidence that occasion comment. Their conclusion depended on inventing their own characterization of NCM (de Queiroz and Poe, 2003, p. 358):

> Under standard parsimony methods and likelihood models that assume no common mechanism, different characters are assumed not to evolve under the same evolutionary processes.

"Assumed not to," they said, rather than "not assumed to." This gave the impression that different characters would be required to have different rates, and the situation de Queiroz and Poe thought reasonable to expect might not have satisfied that condition. But Tuffley and Steel's (1997, p. 597) own characterizations were more lenient:

> By "no common mechanism," we mean that we may choose a different vector of mutation probabilities for each character, rather than requiring all of them to evolve according to a single vector of mutation probabilities, as is usually the case.

And (Tuffley and Steel, 1997, p. 598):

> By allowing a different vector [of rates] *p* for each character, we are allowing different mechanisms to operate at each site, and the characters may be said to evolve with "no common mechanism."

Different rates are allowed, not required, and "we may choose a different vector", not "we must". The rates may be chosen with complete freedom, and this includes the possibility that some of them are alike. Even if observed, what de Queiroz and Poe thought reasonable to expect would not refute *Tuffley and Steel's* NCM. It would seem that de Queiroz and Poe, forced by their preference in methods to use unrealistic models, were determined to create the impression that parsimony would also require an unrealistic model. That their claim itself depended on an unrealistic (i.e. false) assertion about NCM does not seem to have deterred them in the slightest.

[5]This is not to say that de Queiroz and Poe (2001) understood Popper correctly in all respects. They certainly did not, as Siddall (2001) has elegantly pointed out.

de Queiroz and Poe (2003) also objected to my 1983 paper; in fact they had two different objections. One was (de Queiroz and Poe, 2003, p. 362):

> These statements [by Kluge (2001)] are holdovers from an older interpretation (… Farris, 1983) of how cladistic parsimony methods conform with Popper's falsificationist philosophy, an interpretation that we argue is seriously flawed because it rests on a questionable assumption not required by more recent interpretations based on Popper's degree of corroboration (… Farris et al., 2001). Under the older interpretation, characters that are incongruent with a particular phylogenetic hypothesis are viewed as falsifiers of that hypothesis.

Presumably the purpose of this was to lure uninformed readers into believing that my paper did not concern relating parsimony to explanatory power. de Queiroz and Poe (2003, p. 361) obviously did not believe that themselves, for their other objection involved explanation:

> Farris's [(1983)] conclusion rests on the questionable proposition that attributing similarities to inheritance explains those similarities whereas attributing them to homoplasy does not count as an explanation but only as a dismissal. On the contrary, both postulated homologies and homoplasies are explanatory in accounting for the occurrences of character states in taxa. Hypotheses of homoplasy may be unparsimonious and *ad hoc* (under parsimony models), but that does not make them nonexplanatory.

That comment seems to admit of two main possibilities. One is that de Queiroz and Poe sincerely intended to provide legitimate, scientific explanations for one or more homoplasies. In that case my earlier observations would apply, including (Farris, 1983, p. 10; quoted in full above):

> If external evidence favors the interpretation of homoplasy, however, that hypothesis is not *ad hoc*.

Then the homoplasies in question would not be *ad hoc*. But it is only *ad hoc* hypotheses of homoplasy that are minimized when selecting a tree, so that de Queiroz and Poe's comment would not be relevant to my procedure. The other possibility is that in calling homoplasies explanatory, de Queiroz and Poe had in mind some such stratagem as declaring, ''That's a homoplasy. Well, that explains why the character is that way!'' In that case one might reasonably dismiss their comment as effectively meaningless. It seems worth noting, however, that if such ''explanations'' were accepted, there would be no reason not to accept such other ''explanations'' as, ''That character departs in respect Y from the tree. That explains why the character is that way.'' As that would apply equally well for any Y and any tree, accepting such ''explanations'' would undermine the idea of identifying a most explanatory tree *even by maximum likelihood*, for then the characters would always be ''explained'' regardless of the tree. de Queiroz and Poe's comment, then, either meant nothing whatever or else it undercut their own position.

Induction

Rieppel (2003), one of the remaining inductionists, wanted to portray systematic methods as inductive rather than Popperian. For this purpose he characterized ''parsimony'' as symmetrical in confirmation and disconfirmation (Rieppel, 2003, p. 262):

> Among those [trees], we choose the one that is supported or confirmed by the largest number of congruent characters. This most-parsimonious hypothesis symmetrically disconfirms alternative hypotheses to the degree that these are inconsistent with the most-parsimonious hypothesis.

Of course ''largest number of congruent characters'' actually described clique methods, but in any case the connection with induction was supposed to be (p. 262):

> Inductive support works symmetrically, confirming or disconfirming theories or hypotheses to a greater or lesser degree. An empirically confirmed hypothesis A disconfirms a rival hypothesis B to the degree to which B is inconsistent with A. So if x confirms hypothesis A, y confirms hypothesis B, and if x carries a greater evidentiary weight than y, then A is confirmed and B is symmetrically disconfirmed. In contrast, Popperian falsification works asymmetrically: If it occurs (if it is accepted that it has occurred), it is conclusive.

But ''is conclusive'' is not a contrast to ''confirming A disconfirms B'', it is simply a change of subject, and in any event Popper's approach is hardly limited to conclusive cases. Indeed, as Popper (1959, p. 406) explained:[6]

> If h is confirmed or corroborated or supported by e so that $C(h, e) > 0$, then (*a*) non-h is always undermined by e, *i.e.* $C(\text{non-}h, e) < 0$, and (*b*) h is always undermined by non-e, *i.e.* $C(h, \text{non-}e) < 0$.

This always holds, even when e is inconclusive. Popper's corroboration exhibits exactly the kind of symmetry that Rieppel claimed as an exclusive feature of induction.

In another attempt to separate parsimony from Popper, Rieppel (2003, p. 263) contended:

> The meaning of ''explanatory power'' as used by Farris (1983) is not coextensive with the meaning as Popper used this term.

By that he meant, as it turned out three pages later (Rieppel, 2003, p. 266; brackets in original):

> For Popper, the class of (negated) observation statements a theory entails constitutes its explanatory power (Laudan and Leplin, 2002; note the difference from Farris's, [1983] use of the term). The empirical content of a theory increases with the increasing number of (negated) observation statements it entails.

In Rieppel's conception, then, my usage of ''explanatory power'' did not agree with Popper's because

[6] In the interests of readability, I have changed Popper's symbols for hypothesis and evidence in this passage.

Laudan and Leplin—who must have been selected for this trait—used "explanatory power" for what Popper called empirical content. That involved at best a remarkable laxity of usage, as can be seen from one of Popper's (1963, p. 391) comments:

> It can now be shown quite simply that the maximum degree of the explanatory power of a theory, or of the severity of its tests, depends upon the (informative or empirical) content of the theory.

While explanatory power is certainly related to empirical content, they are not the same concept, any more than "speed" is the same concept as "speed limit". But even if someone else did manage to confuse the two, that is no criticism of my position.

Rieppel (2003, p. 268) also tried to create the impression that Popper's "formalisms" could not be used in science at all:

> The recent systematics literature (… Farris et al., 2001…) documents extensive use of Popper's formalisms for degree of corroboration, severity of test, etc. The concept of probability used in these formalisms is that of logical probability, the formalisms therefore are metalinguistic in nature … In sum, the formalisms of Popper are statements rendered in the language of philosophy, not in the language of science; *e* stands to *h* not in the relation of an observation to justified belief (hypothesis) but in the relation of logical entailment.

So, one would think, Popper would not assess hypotheses on the basis of observation. But that was merely a deception. The whole point of Popper's corroboration is to assess hypotheses on the basis of observation, that is, the results of tests (Popper, 1959, p. 415):

> As to degree of corroboration, it is nothing but a measure of the degree to which a hypothesis *h* has been tested … it is a measure of the rationality of accepting, tentatively, a problematic guess, knowing that it is a guess—but one that has undergone searching examination.

While his wording was less than explicit, Rieppel's aim in attributing *logical* probabilities to Popper while citing Farris et al. (2001) can only have been to suggest that Farris et al. were mistaken in using *statistical* probabilities to evaluate corroboration, as I did before (Farris, 2000) and above. That suggestion is itself mistaken, as can easily be seen from Popper's (1959) discussion of calculating $P(e)$ and $P(e, h)$ when evaluating C and E for a statistical hypothesis. Note that the probabilities in this passage are written in unrelativized form, that is, with no explicit background, so that $P(e)$ and $P(e, h)$ correspond to $p(e, b)$ and $p(e, hb)$, respectively, in the formulae seen earlier. The b in $P(a, b)$ here does not refer to background knowledge but to a statistical population. Popper (1959, p. 410f) explained:

> Now let h be the statement $P(a, b) = r$ and let e be the statement 'In a sample which has size n and which satisfies the condition b (or which is taken at random from the population b), a is satisfied in $n(r \pm \delta)$ of the instances'. Then we may put, especially for small values of δ, $P(e) \approx 2\delta$. We may even put $P(e) = 2\delta$, for this would mean that we assign equal probabilities—and therefore, the probabilities $1/(n + 1)$—to each of the $n + 1$ proportions, $0/n$, $1/n$,... n/n, with which a property a may occur in a sample of size n... (The equidistribution here described is the one which Laplace assumes in the derivation of his rule of succession. It is adequate for assessing the *absolute* probability, $P(e)$, if e is a *statistical report about a sample*. But it is inadequate for assessing the relative probability $P(e, h)$ of the same report, given a hypothesis h according to which the sample is the product of an n times repeated experiment whose possible results occur with a certain probability. For in this case it is adequate to assume a combinatoric, *i.e.* a Bernoullian [binomial] rather than a Laplacean distribution.)... We therefore find that $P(e,h) - P(e)$, and thus our functions E and C, can only be large if δ is small and n large.

Both the binomial distribution on possible sample frequencies and the discrete uniform distribution $P(e) = 1/(n + 1)$ are obviously ordinary, statistical probabilities, and just as obviously, such probabilities can be used in calculating C and E. It is surprising that Rieppel (2003) was unaware of this, considering that Farris et al. (2001, p. 440) had quoted the same passage.

Finally, Rieppel et al. (2006, p. 186) maintained:

> One [approach], the hypothetico-deductive method, proves a bad fit with phylogenetic systematics because it requires an excessively strong assumption of the relationship that obtains between hypotheses of descent and the available evidence.

The assumption they had in mind was that phylogenetic hypotheses could be deductively assessed on the basis of observed character distributions. In that case homoplasy would have to be impossible, or else (Rieppel et al., 2006, p. 188):

> Only if independent criteria were available to reliably distinguish "true" synapomorphy from homoplasy would a Popperian form of falsificationism be applicable to systematics.

One reviewer of this paper regarded the alleged bad fit as a "fundamental, and if true, devastating claim about the relevance of Popper's work to systematics". That was at best wishful thinking. My (Farris, 2000; Farris et al., 2001) derivation, discussed above, uses Popper's formulae but does not require anything like the assumption Rieppel et al. suggested. To actually show something about the relevance of Popper's work, an argument would need at the least to consider the relevant parts of Popper's work. Among those relevant parts are Popper's use of statistical probabilities, and it has already been seen how Rieppel (2003) avoided considering that subject.

FCT

According to Faith (1992, 1999, 2004, 2006) and coauthors (Faith and Cranston, 1992; Faith and Trueman,

1996, 2001) Popperian corroboration could be assessed by the PTP[7]test (of Faith and Cranston, 1991). The Faith/Cranston/Trueman view (FCT) involved other innovations as well. ''Evidence'' became degree of fit, and ''background knowledge'' became a null model (Faith, 1992, p. 268):

> The *test statement*(*s*) or purported *evidence*, *e*, for the hypothesis is the degree of cladistic structure found, as measured by the length of the mpt [most parsimonious tree]. Thus, the test statement is produced by the application of cladistic parsimony, as a goodness-of-fit criterion, to the observed taxonomic data.
>
> The *background knowledge*, *b*, reflects other provisionally accepted explanations relating to how a given tree length could result from such an analysis. Faith and Cranston (1991, 1992) argue that this general background knowledge can be represented by a null model, in which the characters are free to covary randomly.

While those revisions were presented as Popperian, they seem to have had a further purpose, for they were immediately applied as grounds for objecting to phylogenetic methods (Faith, 1992, p. 267):

> Carpenter [(1992)] emphasizes cladistics' underlying assumption that common ancestry can explain shared features; this assumption therefore might be interpreted as providing our background knowledge [but] … the mpt [most parsimonious tree] is the least, not most. corroborated hypothesis based on this limited view of background knowledge.

That application continued (Faith and Trueman, 2001, p. 342):

> [Corroboration] $p(e, hb)-p(e, b)$ as used in cladistics, (e = data) implies only fit, whereas true corroboration requires consideration of $p(e, b)$ for e equal to fit itself.

Similarly (Faith, 2004, p. 5):

> Suppose we have data, *d* and the assumptions of some phylogenetic inference method, corresponding to a model, *m*. Phylogenetic hypotheses selected to make $p(h, dm)$ large are *ad hoc*, having little empirical content because *h* offers little content beyond the data plus the model.

But neither those criticisms nor the FCT revisions themselves could be justified on legitimate Popperian grounds, as will soon be apparent.

FCT relied on its own ''corroboration'' formula, which Faith (1992) arrived at by replacing Popper's C with $1-p(e, b)$. To accomplish that he simply ignored the denominator of C while maintaining that $p(e, hb)$ would always be unity (Faith, 1992, p. 268):[8]

> The more general expression of corroboration is a function of $p(e, h \& b)-p(e, b)$[*sic*]. However, the first term here is always equal to 1: because *e* is the length of the most parsimonious tree for these data, and *h* implies that a cladistic analysis of the data was carried out, it follows that $p(e, h \& b)$ [*sic*] is equal to the probability of obtaining a most parsimonious tree of the observed length, given a cladistic analysis of the observed data, and this must equal 1.

It will be seen later that such reasoning violates Popper's (1963, p. 288) rule for exclusion of *ad hoc* hypotheses, but Faith ignored that rule just as he did the denominator. Replacing C with $1-p(e, b)$ suited Faith's purposes because then PTP would seem to determine ''corroboration'', since he identified $p(e, b)$ with PTP, that is, with the tail probability from the PTP test (Faith, 1992, p. 268):

> A *low probability*, corresponding to $p(e, b)$, [*sic*] and reflecting by definition a degree of *corroboration* and *boldness*, is found if the corresponding PTP value is low.

Of course, being a tail (cumulative) probability, PTP could hardly be $p(e, b)$ as the latter is a point probability, but Faith (1992) ignored that problem as well. FCT ''corroboration'' thus became what I shall denote Q:

$$Q = 1 - \mathrm{PTP}$$

But this leads at once to still another problem, for Q is obviously a probability—the head probability complementary to the PTP tail probability—and a legitimate measure of corroboration cannot be a probability, as Popper (1983, p. 243) emphasized:

> For all satisfactory definitions—that is to say for all those which are topologically equivalent to [the formula for C above]—the following theorem holds: Degree of corroboration is not a probability; that is to say, it does not satisfy the rules of the calculus of probability.

The same problem would apply to $1-p(e, b)$ itself, which is also a probability, $p(\text{non-}e, b)$. Faith (1992) ignored that problem too, nor did any advocate of FCT ever address it.

But Faith and Trueman (2001) took another approach to justifying Q as ''corroboration''. As they presented it, every step of turning Popper's C into Q followed from their fit assumption, their view that ''evidence'' meant fit. The fit assumption provided their reason for discarding the denominator of C, as will be seen later, and it also furnished their rationale for maintaining that $p(e, hb)$ must always be unity (Faith and Trueman, 2001, p. 334):

> $p(e, hb) = 1$ because *h*, as given, provides the observed level of fit.

They invoked the fit assumption once more to defend identifying cumulative probability PTP with Popper's $p(e, b)$. The latter, they explained, had to be

[7]So called because in 1990 I suggested the name (an acronym for ''permutation tail probability''; cf. Farris, 1996) in a conversation with Faith. Faith and Cranston (1991) thereupon arrived at the PTP test by renaming a test originally introduced by J. Archie in 1985 (see Legendre, 1986, p. 137; cf. Archie, 1989). The method and the name aside, however, the idea that PTP could assess corroboration was entirely Faith's own.

[8]Faith (1992, 1999) always dropped the italics from Popper's *p*.

a cumulative probability, precisely because "evidence" meant level of fit (Faith and Trueman, 2001, p. 334):

> $p(e, b)$ reflects the probability that e^* [from a randomization] will match e. Note e is indicated by the tree length, but, as evidence, it implies that a level of fit has been achieved ("degree of cladistic structure", Faith, 1992); this evidence, therefore, would be matched by any randomization that produced a fit as good or better.

Finally, they appealed to the fit assumption again in order to dispose of a problem created by identifying "corroboration" with $1-p(e, b)$. The problem had been pointed out by Carpenter et al. (1998, p. 107):

> By pretending that $p(e, hb)$ could be frozen at unity "so that the first term can be ignored", Faith (1992: 266) arrived at a formulation [$1-p(e, b)$] that would absurdly make the "corroboration of h" independent of h.

$1-p(e, b)$ would be independent of h because (Farris, 1995, p. 115):

> In Popper's formula, $p(e, b)$ is the probability of the evidence given just the background, without the hypothesis. It therefore makes no sense to say [as Faith (1992) did] that a certain h makes $p(e, b)$ highest (or lowest, for that matter).

But Faith and Trueman (2001, p. 336) explained:[9]

> For the inclusive framework [FCT], $p(e, b)$ varies among hypotheses because evidence is based on fit.

Of course all that would have applied only *if* "evidence" had been fit. To establish that "evidence" *was* fit, Faith and Trueman (2001, p. 336f) offered an example:

> Popper's own writings document how evidence is linked directly to the hypothesis of interest, for example, "In order to find a good test statement e—one which, if true, is highly favorable to h—we must construct a statistical report e…" (Popper, 1959:410). Far from the claim [by Farris (1995)] that "evidence and hypothesis in his [Popper's] formulae were meant to be separate", Popper presents examples where evidence is derived from the hypothesis and argues that such a link is desirable. We conclude that evidence e as goodness-of-fit derived by using tree h provides a sensible Popperian test-statement.

Despite their mention of "examples" this was the only one they presented, and their one example consisted entirely of the sentence fragment quoted here—the ellipsis was theirs. As one might then suspect, the most interesting aspect of their example was what they omitted. That sentence fragment was extracted from Popper's (1959) discussion of statistical hypotheses, of which the part concerning Popper's calculation of $P(e)$ was quoted above in connection with Rieppel's views, and that part does not support Faith and Trueman's conclusions. In their view, recall (Faith and Trueman, 2001, p. 336; underlining added):

> For the inclusive framework, $p(e, b)$ varies among hypotheses because evidence is based on fit.

In that case *Popper's* evidence is certainly not based on fit, for $p(e, b)$ corresponds to $P(e)$ and Popper's $P(e) = 1/(n + 1)$ does not vary among hypotheses. $P(e)$ depends only on the sample size n, not on the hypothesis h, which is specified by the value of r in $P(a, b) = r$. Farris (1995, p. 115; quoted above) and Carpenter et al. (1998, p. 107; quoted above) were thus entirely correct in pointing out that $p(e, b)$ does not depend on h, and that, consequently, $1-p(e, b)$ can hardly measure the corroboration of h. But then $1-p(e, b)$ could hardly be corroboration in any case, being a probability.

The same example refutes the FCT identification of PTP with $p(e, b)$. According to Faith and Trueman (2001), their fit assumption implied that $p(e, b)$ must be a cumulative probability. Yet Popper's discrete uniform distribution $P(e) = 1/(n + 1)$ cannot possibly be a cumulative probability and is plainly a point probability, in which case cumulative probability PTP cannot be identified with $p(e, b)$. Indeed, Faith and Trueman's fit assumption itself is obviously false. On that assumption $p(e, b)$ would be a distribution on degrees of fit. Yet Popper's $P(e)$ is clearly not a distribution on degrees of fit, but is instead a distribution on possible sample proportions, that is, possible data. $P(e)$ assigns the same probability "to each of the $n + 1$ proportions, $0/n$, $1/n$, ... n/n, with which a property a may occur in a sample of size n" (Popper, 1959, p. 411; quoted in full above).

Replacing Popper's C with Q was thus entirely unjustified, and Faith and Trueman's (2001) fit assumption was flatly incompatible with Popper's concept of evidence. But then how did Faith and Trueman conclude just the opposite? Of course they avoided any mention of how Popper calculated $P(e)$, but that was not the only relevant information that they omitted. Consider their argument again and recall that the ellipsis in the quotation was theirs (Faith and Trueman, 2001, p. 336; quoted in full above):

> Popper's own writings document how evidence is linked directly to the hypothesis of interest, for example, "In order to find a good test statement e—one which, if true, is highly favorable to h—we must construct a statistical report e…" (Popper, 1959, p. 410).

What they quoted from Popper referred to identifying a test-statement that, if true, would be highly favorable to h, but they took the quotation to mean that "evidence is linked directly to the hypothesis of interest" and that (same page; quoted in full above):

> Popper presents examples where evidence is derived from the hypothesis.

[9] As will be seen later, Faith and Trueman (2001) based their "inclusive philosophical framework" on their supposition that corroboration could be "decoupled" from falsification—an idea so thoroughly mistaken that it seems best to avoid the term.

They argued, that is, as if "evidence" would be just the favorable-if-true test-statement itself. But in reality evidence need not even be consistent with such a test-statement, and in fact that was just what Faith and Trueman's ellipsis concealed, as is evident from Popper's (1959, p. 410) unabridged comments:

> In order to find a *good* test-statement *e*—one which, if true, is highly favourable to *h*—we must construct a statistical report *e* such that (i) *e* makes *P*(*e*,*h*)—which is Fisher's likelihood of *h* given *e*—large, *i.e.* nearly equal to 1, and such that (ii) *e* makes *P*(*e*) small, *i.e.* nearly equal to 0. Having constructed a test statement *e* of this kind, we must submit *e* itself to empirical tests. (That is to say, we must *try* to find evidence refuting *e*.)

A test-statement that, *if true*, would be highly favorable to *h* is of course a prediction of *h*, in particular one for which *P*(*e*) is much smaller than *P*(*e*, *h*). But making the prediction—formulating the favorable-*if-true* test-statement—does not in itself produce evidence. On the contrary, testing the prediction (and so the hypothesis) requires *empirical* evidence—observation—and a sincere attempt to find observations that *refute* that test-statement.

Faith and Trueman's (2001) attempt to create the impression that "evidence is derived from the hypothesis" thus rested on using ellipsis to conceal the distinction between prediction and empirical observation, so allowing the pretense that "evidence" consisted of the predictions of a hypothesis. Of course that pretense was absurd, but Faith and Trueman found it useful because the idea that "evidence is derived from the hypothesis" was connected to other aspects of FCT. To support his claim that *p*(*e*, *hb*) must be unity, for example, Faith (1992, p. 266) maintained that the "evidence" would follow from *h* and *b*:

> The first term [of *C*] is the probability of *e* given both *h* and *b*, and will be unity when *e* follows from *h* and *b*. In the present context, this will be the case (see below) [p. 268, quoted above], so that the first term can be ignored.

If that were correct, "evidence" *would* be "derived from the hypothesis". With an accurate reading of Popper's discussion, however, such connections instead uncover further weaknesses of FCT. Thus the fact that empirical evidence may refute a prediction of *h*, directly contradicts the FCT supposition that *p*(*e*, *hb*) must always be unity, for as Popper (1983, p. 242) pointed out:

> What about an empirical evidence *e* which falsifies *h* in the presence of *b*? Such an *e* will make *p*(*e*, *hb*) equal to zero.

Faith and Trueman relied on omission again in objecting to what they portrayed as a claim of mine (Faith and Trueman, 2001, p. 336; quoted in full above):

> Far from the claim [by Farris (1995)] that "evidence and hypothesis in his [Popper's] formulae were meant to be separate", Popper presents examples where evidence is derived from the hypothesis and argues that such a link is desirable.

What they did not mention was *Popper's* reason for keeping the hypothesis separate from the evidence, although I had called attention to it (Farris, 1995, p. 115):

> [Faith's (1992, p. 267)] "p(*e*, *h* & *b*) = 1 for all *h*" means that any tree *h* is supposed to make its length *e* certain, but in fact a tree by itself has no such effect. A tree determines a length only in combination with data… [Faith (1992)] has mixed the data into the tree, ignoring Popper's (1963, p. 288) warning, quoted before, that including such *ad hoc* elements only makes nonsense of "corroboration".

This referred to Popper's (1963, p. 288f) discussion of *ad hoc* hypotheses:

> My definition [of corroboration] does not automatically exclude *ad hoc* hypotheses, but it can be shown to give most reasonable results if combined with a rule excluding *ad hoc* hypotheses. [footnote] This rule for the exclusion of *ad hoc* hypotheses may take the following form: the hypothesis *must not repeat* (except in a completely generalized form) the evidence, or any conjunctive component of it. That is to say *x* = 'This swan is white', is not acceptable as a hypothesis to explain the evidence *y* = 'This swan is white' although 'All swans are white' would be acceptable; and no explanation *x* of *y* must be circular in this sense with respect to any (non-redundant) conjunctive component of *y*.

If the data were included in the hypothesis, then the hypothesis would obviously repeat a conjunctive component of the evidence, violating Popper's rule for exclusion of *ad hoc* hypotheses. Popper's example of evidence, 'This swan is white', it may be added, shows no sign of being a level of fit.

Faith and Trueman (2001) did not entirely repeat Faith's (1992) argument, but they did repeat his mistake. They still identified *e* with fit and they still wanted to conclude that (Faith and Trueman, 2001, p. 334):

> *p*(*e*, *hb*) = 1 because *h*, as given, provides the observed level of fit.

How *h* was supposed to provide fit became clear on the next page (Faith and Trueman, 2001, p. 335; underlining added):

> In the inclusive framework, the fit value interpreted as evidence *e* follows from the nominated hypothesis, <u>given the data</u> and the method defining fit.

Of course the fit value would follow from the hypothesis *given the data*! But this would not imply *p*(*e*, *hb*) = 1 unless the data were included in the condition *hb* of that conditional probability, and the FCT background consisted just of the randomization null model. To arrive at *p*(*e*, *hb*) = 1, then, Faith and Trueman evidently planned to include the data in the hypothesis *h* just as Faith (1992) had done, again violating Popper's rule for exclusion of *ad hoc* hypotheses.

Faith and Trueman (2001) never addressed that rule; neither did Faith (2004, 2006). Evidently they could deal with that difficulty only by avoiding any reference to it.

But then much the same applies to the rest of Popper's comments. The sentence fragment Faith and Trueman presented as proof that Popper would agree with their views is actually the beginning of a section that refutes their assumptions that "evidence" would mean fit, that Popper's $p(e, b)$ would depend on the hypothesis, that $p(e, b)$ would be a cumulative probability, that $p(e, hb)$ would always be unity, and that Q could be used in place of C. Faith and Trueman were able to maintain their position only because they avoided mentioning the relevant parts of Popper's discussion.

Tails

Nor was that situation atypical, as will be evident from Faith and Trueman's (2001) further defense of their supposition that cumulative probability PTP could be identified with Popper's $p(e, b)$. Farris et al. (2001) made two observations on that subject. The first involved likelihood (Farris et al., 2001, p. 439):

> Because PTP is a cumulative probability, equating PTP with $p(e, b)$ would mean that $p(e, b)$ must also be a cumulative probability; if so, then $p(e,hb)$ would have to be a cumulative probability as well, because $p(e, hb)$ differs from $p(e, b)$ only in the added condition h. Yet $p(e, hb)$ cannot be a cumulative probability, for it is a likelihood. Maximum likelihood estimation procedures always maximize point probabilities or densities (see Lindgren, 1962), not cumulative probabilities.

The second has already been discussed (Farris et al., 2001, p. 440):

> [In Popper's (1959, p. 410f) discussion of statistical hypotheses, quoted above] $P(e, h)$ is thus the Bernoullian (binomial) probability of obtaining so many a's with sample size n and parametric frequency r. That is a point, not a cumulative, probability and the same is obviously true of the discrete uniform distribution $P(e) = \delta = 1/(n + 1)$.

No advocate of FCT ever addressed the second point, but Faith and Trueman (2001), who had seen Farris et al.'s manuscript before submitting their own paper, did respond to the comment about likelihood—in a way. They avoided any mention of $p(e, hb)$ as likelihood while "restating" the formula for C so that "$p(e, hb)$" was turned into something else (Faith and Trueman, 2001, p. 334):

> Inclusiveness of the corroboration framework can be understood by restating formula 1 for corroboration. Ignoring [the denominator], the corroboration C of an hypothesis h, given data d, fit-as-evidence e, method/model/assumptions m, and background knowledge b, is:
>
> $$C(h, e, b) = p[\mathrm{Fit}(h, d, m), hb] - p[\mathrm{Fit}(h, d, m), b]$$

That is, they gratuitously substituted "Fit (h, d, m)" for Popper's e in the numerator of C. Among possible measures of fit (see their Table 1) they listed the likelihood of the hypothesis, which they wrote as $p(d, hm)$, for the probability of data d given h and "method/model/assumptions" m. Being a likelihood, $p(d, hm)$ would be a point probability, but—on their view—the p in the formula for C instead denoted cumulative probability, so that "$p(e, hb)$" would have been a cumulative probability $K(p(d, hm), hb)$ of a likelihood value.

Of course their restated formula was sheer invention. Nothing in Popper's (1959, 1963, 1972, 1983) discussion resembles Faith and Trueman's (2001) formulation, and their position is immediately seen to be vacuous, as Popper's $p(e, hb)$—that is, $P(e, h)$—is itself Fisher's likelihood (Popper, 1959, p. 410; braces and underlining added):

> { In order to find a *good* test-statement e—one which, if true, is highly favourable to h—we must construct a statistical report e } such that (i) e makes $P(e, h)$—which is Fisher's likelihood of h given e—large, *i.e.* nearly equal to 1, and such that (ii) e makes $P(e)$ small, *i.e.* nearly equal to 0.

Again Faith and Trueman relied on ellipsis. The part within braces is the sentence fragment that they (Faith and Trueman, 2001, p. 336; quoted above) tried to use as evidence that Popper would agree with them, whereas the underlined part refutes their position. They avoided that difficulty by the simple expedient of cutting Popper off in mid-sentence.

Faith and Trueman (2001) employed their elliptical quotation technique still again when discussing some earlier comments that I (Farris, 1995) had made on the cumulative probability issue. As they presented it, I had misidentified PTP as a point probability (Faith and Trueman, 2001, p. 333):

> Farris (1995, p. 107) claimed that "Faith's version of $p(e, b)$ is… the probability of obtaining the observed L [tree length]under the null model…", and that PTP is equated with this quantity. PTP, which is indeed equated with $p(e, b)$, would then correspond to calculating the permutation tail probability as a point probability. However, in all its applications, PTP has never been calculated as a point probability and indeed was defined (Faith and Cranston, 1991, 1992) as a tail probability.

Note that the ellipses were theirs. It will be informative to compare their version with my actual comments (Farris, 1995, p. 107):

> Faith's version of $p(e, b)$ is then $p(L, N)$, the probability of obtaining the observed [tree length] L under the null model N. That Faith equates logical probability [*i.e.* $p(h, b)$] with both PTP and $p(L, N)$ poses a difficulty in itself, for PTP can differ greatly from $p(L, N)$ in value, as will be illustrated later.

That illustration of the difference between $p(L, N)$ and PTP was included in my discussion of a further problem, that Faith (1992) had substituted PTP for $p(L, N)$—*i.e.* for $p(e, b)$—when evaluating "corroboration" (Farris, 1995, p. 109):

> Substituting PTP for $p(L, N)$ is simply bad statistics. PTP is a tail (cumulative) probability, while $p(L, N)$ is the probability of a single value. The distribution from Farris (1991: 89) example ONE illustrates the distinction:

L	4	5	6	7
$p(L,N)$	0.022	0.276	0.690	0.012
PTP	0.022	0.298	0.988	1.000

In view of this example and my observation that "PTP is a tail (cumulative) probability", it is not possible that Faith and Trueman (2001) actually believed that I had calculated PTP as a point probability. Evidently they intended to mislead their readers, and their purpose in this can only have been to conceal the fact—which they never addressed—that cumulative probability PTP cannot sensibly be identified with point probabilities such as Popper's $p(e, b)$ and $p(h, b)$.

I discussed Faith and Trueman's (2001) use of ellipsis in a paper that I presented at the 2001 meeting of the Willi Hennig Society. Faith was also present, and perhaps for that reason he later took a different approach to defending the FCT position on cumulative probabilities, seeming to confront the issue directly (Faith, 2004, p. 6):

> Farris et al. (2001) also argued that Popper used only point probabilities, while a PTP test uses a tail probability. But the discussion of the Neptune example illustrates how such a tail probability is a natural part of any assessment of degree of improbability of evidence that good (or better).

That directness proved illusory, for Faith never addressed Farris et al.'s (2001, p. 239f; quoted above) comments on likelihood and the obvious point probability $P(e)$. Nor did he provide any explanation of how he arrived at his own position (Faith, 2004, p. 4f):[10]

> Popper (1959, p. 415) argues that logical probability can be suggested by statistical material, and that the quantification of probabilities and content statements that make up corroboration may depend on statistical considerations:
>
> apart from those applications of probability theory in which we can measure probabilities in the usual way (with the help of either the assumption of equal probabilities as in dicing or with the help of statistical hypotheses) I see no possibility of attaching numerical values (other than zero or one) to our measures of probability or content (Popper, 1963, p. 397).
>
> This supports the interpretation of the PTP tail probability as an indicator of the $p(e, b)$ value critical to corroboration assessment.

He seems to have drawn his conclusion about PTP completely gratuitously, from a passage that does not mention tail probabilities at all and in fact concerns a different subject.

Faith's (2004) argument from the Neptune example was even more obscure, as he cited Popper (1983) without providing page numbers. Fortunately, Popper (1983) mentioned Neptune in only three passages, beginning with (Popper, 1983, p. 237):

> For example, let e be the first observation of a new planet (Neptune) by J. G. Galle, in a position predicted by Adams and Leverrier, and let h be Newton's theory upon which their prediction was based. Then e certainly supports h—and very strongly so. Yet in spite of this fact e also follows from theories which, like Einstein's, entail non-h (in the presence of b).

The support is strong because (Popper, 1983, p. 247):

> Adams and Leverrier's predictions, which led to the discovery of Neptune, were such a wonderful corroboration of Newton's theory because of the exceeding *improbability* that an as yet unobserved planet would, by sheer accident, be found in that small region of the sky where their calculations had placed it.

That the object was a *planet*—was moving—is the crucial point (Popper, 1983, p. 237):

> Thus James Challis, to whom Adams had given the results of his calculations, actually observed Neptune close to the calculated orbit before Galle. But the star he saw did not seem to move, and he did not think his observation sufficiently significant to compare it with later observations of the same region which would have disclosed its motion. The presence of *some* unknown star of eighth magnitude close to the calculated place, was in itself quite probable on his background knowledge and therefore did not appear significant to him. Only that of a *moving* star, a planet, would have been significant, because unexpected—though not on Adam's calculations.

How was this supposed to lead to PTP? According to Faith (2004, p. 4):

> The improbability involved a tail probability analogous to that of PTP; Neptune was observed some x units from the predicted place, but the evidence statement for Newton's theory was the observation of a planet that close—x units or less (not the observation at exactly x units).

Yet Popper did not say "tail probability", or "x units or less", or even "x units from the predicted place", but that Neptune was observed "in a position predicted by Adams and Leverrier". The improbability that made for strong corroboration, furthermore, had nothing to do with tail probabilities, but arose instead from the fact that Neptune was moving rather than stationary. Faith's conclusion was again purely gratuitous, merely an attempt to put his own words into Popper's mouth.

I called attention to Faith's (2004) gratuitous conclusion of tail probabilities in a paper that I presented at the 2004 meeting of the Societé Française Systematique, in a session at which Faith was also present. Perhaps for that reason, Faith (2006) made no attempt to show by any quotation, or even gratuitous paraphrase, that Popper used tail or cumulative probabilities. In fact, although

[10] Popper's (1963, p. 397) comment actually read "0 or 1" instead of "zero or one".

Faith (2006) still persisted in referring to PTP as assessing corroboration, he never mentioned the cumulative probability issue. It would appear that he could no longer think of any reason for his position.

The situation was no better for the FCT assumption that "evidence" meant fit, Faith's (2004) defense of which was also based on the Neptune example (Faith, 2004, p. 4):

> The evidence for Newton's theory was a measure of the degree of fit of observations to the hypothesis and corroboration depended on a judgment that this evidence was improbable without the hypothesis.

This was still another gratuitous interpretation, for Popper (1983, p. 237; quoted in full above) said that the evidence was the position of a new planet:

> For example, let e be the first observation of a new planet (Neptune) by J. G. Galle, in a position predicted by Adams and Leverrier, and let h be Newton's theory upon which their prediction was based.

Once more Faith simply substituted his own words for Popper's, as indeed became even more obvious in the 2006 version of his comments (Faith, 2006, p. 555):

> In one such example, the hypothesis was Newton's theory and the *evidence* was the observation of a new planet, Neptune, in a position close to [*sic*] that predicted by the hypothesis. Thus, the evidence was given by *goodness-of-fit* of observations to the hypothesis.

Faith's argument amounted to a play on words. He found an example in which the evidence *did fit* the hypothesis and portrayed that as meaning that "evidence" *was* fit! But it would have been harder for plays on words to obscure the information to which Farris et al. (2001, p. 440) called attention:

> Popper's (1959, p. 410f; quoted above)] $P(e)$ and $P(e, h)$ are simply distributions on the number of a's in a sample of n independent observations, not on any variable that could be regarded as a measure of fit. Identifying evidence with fit, in fact, directly violates a rule that Popper (1972, p. 288) discussed while emphasizing the importance of avoiding *ad hoc* hypotheses: [Popper's rule for exclusion of *ad hoc* hypotheses, quoted above].

Accordingly, no advocate of FCT ever addressed those points. Again, FCT could be maintained only by avoiding mention of the relevant parts of Popper's discussion.

Unity

Faith and Trueman (2001) also tried to defend their assumption that $p(e, hb)$ must be unity. Farris et al. (2001, p. 439) had refuted that assumption by quoting Popper (1983, p. 242):

> What about an empirical evidence e which falsifies h in the presence of b? Such an e will make $p(e, hb)$ equal to zero.

Faith and Trueman's response consisted of never mentioning Popper's actual comment and substituting their own version (Faith and Trueman, 2001, p. 335):

> Popper (1959) sees low corroboration as corresponding to some degree of falsification of h exactly when e is "negative" evidence interpretable as falsifying h. On such occasions, $p(e, hb)$ is < 0, but $p(e, b)$ typically is positive, and we have negative values for corroboration given by $p(e, hb) - p(e, b)$.

Their version was complete nonsense, as no probability can be < 0. It seems remarkable that the author of a significance test would have been unaware of that fact.

Fixing $p(e, hb)$ artificially at unity leads to a difficulty that I have pointed out before (Farris, 1995). As PTP cannot exceed unity, setting $p(e, hb) = 1$ in FCT "corroboration" $Q = 1-\text{PTP}$ means that Q cannot be negative. But according to Popper (1983, p. 241):

> If e supports h (given the background knowledge b) then $C(h, e, b)$ is positive. If e undermines h (so that non-e supports h) then $C(h, e, b)$ is negative. If e does neither, so that it is *independent* of h in the presence of b, then $C(h, e, b)$ equals zero.

If interpreted as if it were Popper's C, then, Q could never have indicated unfavorable evidence. Faith and Trueman (2001) did not acknowledge this problem directly, but they did propose what might have seemed to be a solution. Their fit assumption, they explained, would imply that evidence would always be favorable, with the added benefit that the denominator of C could be dispensed with (Faith and Trueman, 2001, p. 335):

> When the evidence considered for phylogenetic trees is always positive (as it is for goodness-of-fit measures: trees differ only in how good is the goodness-of-fit), the standardization function [denominator] in formula 1 [for C] can be ignored.

This would not have been limited to trees. The same would have applied whenever "evidence" could be equated to fit, and—on the FCT view—that would have included every field from which Popper had ever drawn an example. Now it was no longer a defect that fixing $p(e, hb)$ at unity would prevent recognition of unfavorable evidence, because there was no longer such a thing as unfavorable evidence!

Faith and Trueman attributed great importance to that idea. In their conception, doing away with unfavorable evidence—and so with falsification—allowed a "decoupled" interpretation that would provide the basis for a whole new approach, an "inclusive philosophical framework" (Faith and Trueman, 2001, p. 331):[11]

> We defend and expand our earlier proposal for an inclusive philosophical framework for phylogenetics, based on an interpretation of Popperian corroboration that is decoupled from the popular falsificationist interpretation of Popperian philosophy.

[11] "Expand our earlier proposal for an inclusive philosophical framework" was a pure example of revising history. No previous FCT paper had mentioned an "inclusive framework".

In the process it decoupled their approach from Popper's (1983, p. 188):

> Our tests are *attempted refutations*; [they] are designed—*designed in the light of some competing hypothesis*—with the aim of refuting, if possible, the theory which we wish to test.

This is entirely incompatible with the supposition that evidence can only be favorable, for obviously there can be no sincere attempt at refutation unless it is logically possible to obtain evidence against the hypothesis being tested. The same conflict means that Faith and Trueman's idea, considered as a description of *Popperian* corroboration, was patently false. If evidence were always favorable, Popper would never have regarded any hypothesis as refuted, whereas (Popper, 1963, p. 242):

> An example is the marvelous theory of Bohr, Kramers and Slater of 1924 which, as an intellectual achievement, might even rank with Bohr's theory of the hydrogen atom of 1913. Yet unfortunately it was almost at once refuted by the facts—by the coincidence experiments of Bothe and Geiger.

And (same page):

> Even Newton's theory was in the end refuted.

Further (Popper, 1963, p. 239):

> An excellent recent example is the rejection, in atomic theory, of the law of parity; another is the rejection of the law of commutation for conjugate variables.

Indeed (Popper, 1959, p. 131):

> The unequivocal *negative* result which Kepler reached by the falsification of his circle hypothesis was in fact his first real success.

But even beyond being obviously false, Faith and Trueman's (2001) position was literally paradoxical. As was seen earlier, evidence favorable to h must be unfavorable to non-h, and conversely (Popper, 1959, p. 406):

> If h is confirmed or corroborated or supported by e so that $C(h, e) > 0$, then (a) non-h is always undermined by e, *i.e.* $C(\text{non-}h, e) < 0$, and (b) h is always undermined by non-e, *i.e.* $C(h, \text{non-}e) < 0$.

In that case it is clearly impossible for evidence to be always favorable.

Although that property of Popperian corroboration had been pointed out before, by Carpenter et al. (1998),[12] Faith and Trueman (2001) never addressed the paradoxical nature of their position. Nor did Faith (2004, 2006),[13] but by then he had already abandoned the decoupled interpretation. This emerged in Faith's response to Rieppel (2003), who at least was aware that Faith and Trueman's (2001) position was inconsistent with Popper's (Rieppel, 2003, p. 269; braces added):

> Faith and Trueman (2001:331) escaped from large parts of what I have explained because they explicitly decoupled corroboration from "the popular falsificationist interpretation of Popperian philosophy" (rendering it Popper*: Faith, 1999). Thus, the meaning with which Popper used the concept of degree of corroboration cannot be coextensive with the meaning bestowed on that concept by Faith and Trueman (2001; see remarks above on semantic analysis)… {Popper linked corroborability to testability (Popper, 1983, p. 245) and stated that "scientific tests are always attempted *refutations*" (Popper, 1983, p. 243). For him, corroboration was thus embedded in a falsificationist context.}

Faith (2004) did not mention the first part of Rieppel's discussion—the part that described Faith and Trueman's (2001) views—but cited only the comment within the braces. To that comment he replied (Faith, 2004, p. 7):

> In response, I note first that the focus of the inclusive framework on corroboration does not deny falsification. But more importantly, it does link corroboration emphatically to tests as attempted refutations.

Now the "inclusive framework" did *not* involve decoupling corroboration from falsification! While hardly presented as such, this was in fact a retraction. Or rather it was an erasure, for Faith (2004, 2006) never mentioned the decoupled interpretation or Faith and Trueman's (2001) idea that evidence would always be favorable. Nor, accordingly, did Faith (2004, 2006) any longer mention the underlying reason for those ill-considered proposals, his (Faith, 1992) erroneous supposition that $p(e, hb)$ must always be unity.

Admitting that $p(e, hb)$ can vary led to new problems for FCT. Like Faith and Trueman (2001, p. 335; quoted above) Faith (2004, p. 6) still wanted "evidence" to mean fit:

> In the inclusive framework, evidence e typically follows from the nominated hypothesis, as a statement of that hypothesis' fit to some observed data.

As before, *if e* followed from h, then $p(e, hb)$ would be unity. Consequently, while he certainly did not admit it explicitly, by abandoning his claim that $p(e, hb)$ must be unity, Faith (2004) contradicted his view of "evidence".

[12] Carpenter et al. (1998) were commenting on the position of Faith and Trueman (1996), which was just the opposite of Faith and Trueman's (2001) in this respect, but just as paradoxical. According to Faith and Trueman (1996, p. 582), "In fact, both monophyly and nonmonophyly hypotheses could be falsified." When the observed value for D falls in the middle of the distribution of difference values (Fig. 2), both hypotheses would be falsified."

[13] Except that Faith (2004, p. 9) tried to change history: "In de Queiroz and Poe's criticisms of PTP as corroboration assessment, low corroboration of h is claimed to necessarily imply high corroboration of a not-h hypothesis. However, both hypotheses could have low Popperian corroboration, based on the current evidence (Faith and Trueman, 1998 [*sic!*—it should be 1996] present a simple example)."

Faith (2004, p. 2) also still wanted to interpret PTP as assessing corroboration:

> PTP tests (Faith, 1990; Faith and Cranston, 1991) provide one example of corroboration assessment, in using background knowledge based on random character covariation and evidence based on fit from application of cladistic parsimony.

Of course that idea rested on misidentifying PTP with $p(e, b)$, but it also involved neglecting $p(e, hb)$. In Faith's (1992, p. 266; quoted in full above) original argument, this was supposed to be justified by the claim that $p(e, hb)$ would always be unity, "so that the first term can be ignored". But as Faith (2004, 2006) no longer contended that $p(e, hb)$ would always be unity, some new reason for ignoring $p(e, hb)$ would now have been required. Yet he suggested no such reason, and in fact he did not even mention the issue. It seems that one was now supposed to ignore Popper's $p(e, hb)$ purely on faith.[14]

Background knowledge

Because PTP is calculated under a randomization null model, identifying PTP with $p(e, b)$ would require identifying that null model with background knowledge b, and accordingly Faith (1992, p. 268) did so:

> The *background knowledge*, b, reflects other provisionally accepted explanations relating to how a given tree length could result from such an analysis. Faith and Cranston (1991, 1992) argue that this general background knowledge can be represented by a null model, in which the characters are free to covary randomly. The corresponding null hypothesis is that the observed test statement (degree of cladistic structure) [tree length] could have been produced easily under only the assumptions of the null model.

Faith (1992) modeled his "provisionally accepted" on Popper's acceptance requirement for background theories, discussed above, but apparently he did not realize the implications of that requirement. After I pointed out (Farris, 1995, p. 113) that the randomization null model is hardly what anyone could consider an accepted theory, "provisionally accepted" was dropped from later FCT formulations (Faith and Trueman, 2001, p. 333):

> Background knowledge b presents other ways to account for seemingly positive evidence-as-fit.

Similarly (Faith, 2004, p. 3):

> Background knowledge is not the method's assumed way of explaining the data (e.g. descent with modification), but the potential multitude of other possible explanations for our observations.

As one might expect from this, Faith and Trueman (2001) never addressed Popper's acceptance requirement. Neither did Faith (2004, 2006), except that he did mention the subject in just one passage. There he defended his choice of background by citing other discussions (Faith, 2004, p. 6):

> Farris et al. (2001) wrongly assumed that background knowledge must be only well-corroborated theories (with the intended consequence that the null model as used in PTP would be disallowed). This error has been countered in the clarifications and examples above (see also Faith and Trueman, 2001).

In fact those other discussions did not exist.

Naturally Faith would have been aware that the PTP null hypothesis is rejected at a high significance level in almost all real cases, but he wanted to retain the FCT view of background knowledge nonetheless. As a way to keep the null model in b, he resorted to the idea that the null *model*—but not the null *hypothesis*—would be in the background (Faith, 2004, p. 6):

> de Queiroz and Poe (2001) argued that the PTP test tests a null hypothesis, and so this null hypothesis cannot be part of background knowledge, which, according to Popper, is not tested during corroboration assessment. Faith and Trueman (2001) pointed out that, aside from confusing a null model (which is part of b) with a null hypothesis (which is not), the word test in the PTP test is not intended as the test (production of an evidence statement) of Popperian corroboration.

But this put the wrong quantity in the background. PTP is the Type I error rate (α value) for rejecting the null hypothesis (see Farris, 1991), which is to say that PTP is the tail probability for the observed tree length (which Faith called e) given the null hypothesis. Identifying PTP with $p(e, b)$—as Faith wished—would then require, if anything, identifying b with the null *hypothesis*. Even if it were otherwise sensible, furthermore, putting the null model—but not the null hypothesis—in the background still would not save the background from rejection. As the distribution of tree lengths used in Faith's (1992, p. 286; quoted above) null hypothesis was derived from the random character covariation of his null model, rejecting that null hypothesis would logically mean rejecting the null model as well. If the null hypothesis/model is rejected, finally, it scarcely matters whether the rejection is called "production of an evidence statement". As Faith's proposed distinction thus seems to contribute nothing but confusion, I will ignore it.

Rejection is not the only impediment to regarding the null model as background knowledge. There is a further, distinctive drawback (Farris, 2000, p. 387):

> Suppose that the background b included the randomization model R. Then, since (according to R) data matrix e would be entirely independent of tree h, the probability of e given both h and b would be just the probability of e given b. That is, $p(e, hb)$

[14] In a presentation at the 2004 meeting of the Societé Française Systematique, Faith showed Popper's formula for corroboration in what is often called a wardslide—a slide with a black background, the first term $p(e, hb)$ in nearly invisible dark red, and the rest of the formula in brilliant white. He explained that one should concentrate on $p(e, b)$ and think of $p(e, hb)$ as "just going away".

> would equal $p(e, b)$, so that $C(h, e, b)$ would be identically 0. Thus there would be no corroboration: given R, the "evidence" e would be irrelevant to [independent of] h. This should hardly be surprising; after all, the premise of R is that there is no connection between characters and phylogeny.

$E(h, e, b)$ would be identically 0 for the same reason. This and related difficulties have been pointed out repeatedly (Farris, 1995; Carpenter et al., 1998; Farris et al., 2001), but no FCT advocate ever addressed the problem. Faith and Trueman (2001), however, seemed to have a solution—in the same sense that their decoupled interpretation was a solution. They maintained that background knowledge should be chosen to maximize $p(e, b)$ (Faith and Trueman, 2001, p. 340):

> We must try through nomination of background knowledge to make $p(e, b)$ high.

This would generally have minimized corroboration, and it would always have done so on their view, in which "corroboration" was just $1-p(e, b)$. Thus (Faith and Trueman, 2001, p. 341):

> Corroboration assessment is not a search for the right combination of evidence and background knowledge to achieve a low probability [$p(e, b)$]; it is about challenging apparently good evidence by searching for background knowledge that might imply that the evidence was probable anyway.

According to Faith and Trueman, then, it would have been *desirable* to pick the background to eliminate corroboration whenever possible—and it would always have been possible for them, as they were willing to include even rejected null models in background "knowledge"!

But then how did Faith and Trueman justify picking the background to minimize corroboration? As they presented it, the idea was Popper's (Faith and Trueman, 2001, p. 340):

> Popper advocates varying background knowledge, given that any improbability of the evidence is to be achieved despite our best efforts to find background knowledge that shows e to be probable even without h (1963:238; his italics):
>
> > While discussing a problem we always accept (if only temporarily) all kinds of things as *unproblematic*: they constitute for the time being, and for the discussion of this particular problem, what I call *background knowledge*.

Further (1983:188; his italics),

> this background knowledge is usually *varied* by us during the tests, which tends to neutralize mistakes that might be involved in it.

This was only another gratuitous interpretation. Neither of those comments of Popper's mentions minimizing corroboration. On the contrary, according to Popper (1963, p. 288; underlining added):

> The total evidence e is to be partitioned into [evidence part] y and [background part] z; and y and z should be so chosen as to give $C(x, y, z)$ the highest value possible for [hypothesis] x, on the available total evidence.

Obviously the background is *not* chosen to minimize corroboration.

Faith (2004) still maintained that background knowledge should be chosen to minimize corroboration, and he adopted two premises that—if true—would have jointly supported that position. The first, as noted earlier, was that background knowledge would comprise "other possible explanations" (Faith, 2004, p. 3; quoted in full above). The second was that other possible explanations, if found, would preclude corroboration (Faith, 2004, p. 3):

> Suppose that some apparent positive evidence for an hypothesis has been put forward. To judge how well that evidence supports the hypothesis, we can try to explain that evidence away, that is, account for it by possible explanations other than the hypothesis of interest. If, and only if, we fail we can say that the hypothesis has gained Popperian corroboration from that evidence.

Both of Faith's premises were mistaken. The error embodied in the second, in fact, was the point of Popper's (1983, p. 237; underlining added) first comment on Neptune:

> For example, let e be the first observation of a new planet (Neptune) by J. G. Galle, in a position predicted by Adams and Leverrier, and let h be Newton's theory upon which their prediction was based. Then e certainly supports h—and very strongly so. Yet in spite of this fact e also follows from theories which, like Einstein's, entail non-h (in the presence of b).

Faith (2004) again relied on omission. While he cited the Neptune example himself, he never mentioned the underlined sentence. Indeed, Faith (2006) expanded on that omission, going on to claim that same example as support for his position. Again using the idea that strong corroboration requires low $p(e, b)$, he argued (Faith, 2006, p. 555):

> In systematics, this improbability [low $p(e, b)$] can be interpreted to mean that it is difficult to explain-away an achieved degree of fit by other factors (background knowledge). Thus, improbable evidence is evidence that cannot easily be accounted for by some other explanation. This twofold claim as a basis for the inclusive framework—that evidence can be goodness-of-fit, and that improbability of evidence reflects the failure of possible alternative explanations for the evidence—is well supported by Popper's own examples. In one such example, the hypothesis was Newton's theory and the *evidence* was the observation of a new planet, Neptune, in a position close to that predicted by the hypothesis.

It seems noteworthy that Faith (2006) gave no citation for Popper's (1983, p. 237; quoted above) discussion of Neptune, so providing his readers with no means of finding Popper's actual comments. Popper certainly did regard the observed position of Neptune as improbable given the background alone, but this does *not* mean "failure of possible alternative explanations for the evidence". That evidence is also explained

perfectly well by Einstein's theory—as Popper noted in the underlined sentence.

Faith (2006, p. 555) invoked the same example to defend his supposition that background knowledge would consist of "other possible explanations":

> This corroboration [of Newton's theory] therefore depended on a judgment that the evidence, given by a measure of the degree of fit [!] of observations to the hypothesis, was improbable given only background knowledge concerning other possible explanations, including elements of chance.

Like many of his conclusions, this one was purely gratuitous. Popper's (1983, p. 237, 247; quoted above) discussion of Neptune includes no mention whatever of the idea that background knowledge consists of "other possible explanations", and in fact background knowledge *cannot* comprise "other possible explanations". Any sincere effort to find other explanations would have to consider theories other than the hypothesis of interest *h*. When there are such theories (Popper, 1983, p. 188):

> We always try to discover how we might arrange for *crucial tests* between the new hypothesis under investigation—the one we are trying to test—and some others. This is a consequence of the fact that our tests are *attempted refutations*; that they are designed—*designed in the light of some competing hypothesis*—with the aim of refuting, if possible, the theory which we wish to test. And we always try, in a crucial test, to make the background knowledge play exactly the same part—so far as this is possible—with respect to each of the hypotheses between which we are trying to force a decision by the crucial test.

The *test* should be designed to refute *h* if possible. But the background knowledge should be neutral in the choice between *h* and competing theories—explanations—and in that case it can hardly consist of the other explanations. As Popper (1983, p. 236) emphasized:

> It is important to realize that *b* must be consistent with *h*; thus, should we, before considering and testing *h*, accept some theory *h*′, which, together with the rest of our background knowledge is inconsistent with *h*, then we should have to exclude *h*′ from the background knowledge *b*.

Actually Faith (2004, p. 6) was aware of that; he quoted part of the same passage, as did Faith and Trueman (2001, p. 335). Yet they still contended that they could include their null model in background knowledge under "other explanations" (Faith and Trueman, 2001, p. 335):

> Background knowledge, *b* (and the corresponding null hypothesis), does not assume that *h* is false; rather, it assumes that the observed evidence may have other explanations than having been generated under a true *h*.

But this was mere double-talk, for it was a self-contradiction. If, as they said, "generated under a true *h*" would be an explanation of the observed evidence, then $E(h, e, b)$ would be positive, and for a good explanation $E(h, e, b)$ would be close to unity, its maximum value. But if the null model were in the background, $E(h, e, b)$ would instead be zero, as was seen above.

The idea that background knowledge would consist of "other explanations" was never based on any comment of Popper's but was instead entirely Faith's own creation, an outgrowth of his insistence that PTP would assess corroboration. PTP began as an ordinary significance test, in which the null hypothesis of randomness might be rejected in favor of the alternative hypothesis of structured data. But when Faith (1992) tried to turn Popper's $p(e, b)$ into PTP, he had to identify the null hypothesis ("model") with *b*. As that mistake put one of the competing hypotheses in "background knowledge", Faith then had to pretend that background knowledge consisted of "other provisionally accepted explanations"! Faith's need to defend his ill-founded identification of *b* with his null model also led to the other claims reviewed above—that the null model is not really rejected even when the null hypothesis is rejected; that rejected models may nonetheless be "background knowledge"; and that eliminating corroboration is desirable! It might have been more productive if Faith had tried some other way to relate statistics to corroboration.

Corroboration

It is not difficult to relate statistics to corroboration (Farris, 2000; Farris et al., 2001), as was seen at the beginning of this paper, but that solution was unacceptable to advocates of FCT. Faith and Trueman responded much as has been seen before, by not mentioning my actual comments and presenting their own version (Faith and Trueman, 2001, p. 337):

> Suppose we were to equate corroboration with Fit_2 (Table 1) $[p(d, hm) - p(d, m)]$, which has the same basic form as the corroboration formulae. For *any* specific fit criterion and any data, $p(e, hb) = p(d, hm)$ will be high for the best-fit tree and $p(e, b) = p(d, m)$ will be constant over all *h* and ignored. Any fit criterion could create "corroborated" hypotheses under this scheme; all methods could gain this supposed Popperian justification.

No, only methods based on accepted—realistic—background theories would be justified, and there are very few of those. Again Faith and Trueman's argument was based only on avoiding any mention of relevant parts of Popper's discussion and my own.

Any derivation that led to ordinary phylogenetic methods would have been objectionable to advocates of FCT, as the very use of data as evidence would, they said, preclude "true" corroboration (Faith and Trueman, 2001, p. 342):

> $p(e, hb) - p(e, b)$ as used in cladistics, (e = data) implies only fit, whereas true corroboration requires consideration of $p(e, b)$ for e equal to fit itself.

Or alternatively, as Faith (2004, p. 10) maintained:

> However, $p(data, hb)$ will always be less than $p(data, b)$ because the more general model can always fit the data at least as well (remembering that good fit means high probability, p). Thus, corroboration with data as evidence would appear to be zero or negative.

As seen earlier, in his discussion of statistical hypotheses, Popper (1959, p. 410f; quoted above) calculated C using a sample proportion—data—as evidence. Faith and Trueman's (2001) position would thus have implied that Popper's own example of corroboration would not be "true" corroboration! As for Faith's (2004) version, recall an illustration given by Farris et al. (2001, p. 441) but never mention by Faith. In the situation of Popper's discussion, take $r = 0.5$, $n = 1000$, and the observed count of property a to be 500. Applying the formulae, C is then $+0.9238$ so that Faith's claim is obviously false.

Faith (2004) also contended—much like Rieppel (2003, p. 268; quoted above)—that the probabilities used to calculate corroboration could not be probabilities *of truth*, a category that would include the statistical probabilities used in my (Farris, 2000; Farris et al., 2001) derivation. Employing such probabilities, he warned, would lead to verificationism, and that peril could only be avoided by using *logical* probabilities instead. To show why probabilities of truth would be unsuitable, Faith (2004, p. 5) began with Bayes' Theorem:

> These logical probabilities satisfy the usual probability calculus, including Bayes' Theorem [$p(h, eb) = p(e, hb)p(h, b)/p(e, b)$]. Given that corroboration implies that $p(e, hb)$ is greater than $p(e, b)$, Bayes' Theorem implies that corroboration also is reflected in the magnitude of $p(h, eb)/p(h, b)$.

As support for that conclusion he provided a quotation from Popper (Faith, 2004, p. 5):

> According to Popper:
>
> > 'the logical probability of a hypothesis h, relative to the evidence e, also increases with the absolute improbability of e' (Popper, 1983, p. 248; changed to his usual notation).

Faith (2004, p. 5) then argued:

> This greater probability of h as a consequence of corroboration raises important issues, given that Popper attempts to avoid the 'verificationist' process of seeing accumulated evidence for an hypothesis as indicating a greater probability of its truth. Popper (1983) explains that the greater probability associated with corroboration is a statement only about that portion of the content of h accounted for by the evidence. The consequent high $p(h, eb)$ is not interpretable as a high probability that h is true, it only reflects a change in content of h (see also Faith and Trueman, 2001). That argument countering any charge of verificationism depends on logical probabilities. If the same corroboration formulae were to use probabilities that are interpretable as probabilities of truth, then there would be verificationism in the consequent higher value for $p(h, eb)$.

But obviously Popper saw nothing wrong in using statistical probabilities to calculate corroboration, for he did so himself in his discussion of statistical hypotheses (Popper, 1959, p. 410f; quoted above). To create a different impression, Faith again relied on omission, leaving out the part of Popper's explanation, here underlined, that did not suit FCT purposes. Unlike Faith's version, Popper's (1983, p, 248; underlining added) own comment made provision for the case in which the hypothesis has zero probability:[15]

> The logical probability of a hypothesis h, relative to the evidence e, also increases with the absolute improbability of e, provided the absolute logical probability of h was not zero, and that e was derivable or predictable by h.

That case is of particular importance because (Popper, 1983, p. 243):

> Interesting scientific theories have always a negligible (if not zero) probability—including those which are at present generally accepted.

But when $p(h, b) = 0$ corroboration does *not* increase the probability of h, nor is it true that "corroboration also is reflected in the magnitude of $p(h, eb)/p(h, b)$". It is readily seen from Bayes' theorem

$$p(h, eb) = p(e, hb)p(h, b)/p(e, b),$$

that if $p(h, b) = 0$ then $p(h, eb) = 0$, and then $p(h, eb) = p(h, b)$ regardless of the value of C. For example, if

$$p(h, b) = 0, p(e, b) = 0.001, p(e, hb) = 0.1,$$

$p(h, eb)/p(h, b) = 0/0$ is not even well defined, and even though corroboration $C(h, e, b) = 0.98$ is strong, $p(h, eb) = p(h, b)$. It is this characteristic of scientifically interesting theories that is the reason why Popper did not see "accumulated evidence for an hypothesis as indicating a greater probability". The reason has nothing to do with the idea that logical probabilities are not probabilities of truth, and in fact that idea was a figment of Faith's imagination, as Popper (1983, p. 243) made clear:

> Thus, I do not deny that the logical interpretation of probability, or the probability of statements, may be said to give the degree of probability, or likelihood, or chance, of a statement to be true.

Faith (2004) completely misrepresented Popper's way of avoiding verificationism. Popper's approach is not based on attributing some peculiar meaning to "probability", but instead on the realization that high probability is a poor guide for selecting hypotheses (Popper, 1959, p. 399):

[15] I have changed Popper's symbols to h and e to ease comparison with Faith's (2004, p. 5) partial quotation.

> *Science does not aim, primarily, at high probabilities. It aims at a high informative content, well backed by experience. But a hypothesis may be very probable simply because it tells us nothing, or very little.* A high degree of probability is therefore not an indication of goodness.

And (Popper, 1963, p. 248):

> While the verificationists or inductivists in vain try to show that scientific beliefs can be justified, or, at least, established as probable (and so encourage, by their failure, the retreat into irrationalism), we of the other group have found that we do not even want a highly probable theory.

Faith should have known that, since he had once quoted the first of those passages himself (Faith, 1992, p. 266). But there is another connection between verificationism and Faith's (2004) discussion. It can be seen from Popper's explanation of the drawbacks of a conceivable requirement (demand) for evidence e to be regarded as supporting evidence for hypothesis h (Popper, 1983, p. 236f):

> One might be inclined to interpret the remark in question by the demand 'non-e follows from non-h (in the presence of b)'. But this would amount to another form of verificationism: it would make e and h equivalent (in the presence of b), and would thus allow us to verify h by observation, that is, by observing that e is true. But quite apart from any hostility to verificationism, it is altogether implausible to demand that non-e would follow from non-h (and b). For let us assume that e is an event which supports h—something predicted by h, and something nobody would ever have considered without h. For example, let e be the first observation of a new planet (Neptune) by J. G. Galle, in a position predicted by Adams and Leverrier, and let h be Newton's theory upon which their prediction was based. Then e certainly supports h—and very strongly so. Yet in spite of this fact e also follows from theories which, like Einstein's, entail non-h (in the presence of b).

The last part of that passage, as was seen earlier, refutes Faith's contention that finding possible explanations other than h would preclude corroboration of h (Faith, 2004, p. 5):

> Suppose that some apparent positive evidence for an hypothesis has been put forward. To judge how well that evidence supports the hypothesis, we can try to explain that evidence away, that is, account for it by possible explanations other than the hypothesis of interest. If, and only if, we fail we can say that the hypothesis has gained Popperian corroboration from that evidence.

The more extensive quotation from Popper shows the flaws of Faith's position in a broader context. Faith was willing to admit corroboration of h only if no theory other than h could explain e, but in that case non-e would follow from non-h, so that Faith's version of "Popperian corroboration" would, as Popper put it, "amount to another form of verificationism".

Building on his thoroughly unPopperian idea that hypotheses would be selected to maximize $p(h, eb)$, Faith (2004, p. 5) arrived, he said, at another fault of using data as evidence:

> Suppose we have data, d and the assumptions of some phylogenetic inference method, corresponding to a model, m. Phylogenetic hypotheses selected to make $p(h, dm)$ large are *ad hoc*, having little empirical content because h offers little content beyond the data plus the model.

Parts of this argument were accurate. *Ad hoc* hypotheses do have little content, and the content of h does decrease as the probability of h increases (Popper, 1983, p. 241):

> The maximum value which $C(h, e, b)$ can attain is equal to $1-p(h, b)$ and therefore equal to the content of h relative to b, or its degree of testability.

Thus maximizing $p(h, eb)$ could easily lead to an *ad hoc* hypothesis. One could pick a useless hypothesis such as h_1 = "tree t is most parsimonious for data e", which (barring errors in calculating t) would have $p(h_1, eb) = 1$ and no content whatever beyond that of the data. But there was nothing accurate in Faith's contention that Popperians would choose hypotheses to maximize $p(h, eb)$. Phylogeneticists instead identify the most parsimonious tree t for the data e and then provisionally conjecture h_2 = "t is the true phylogeny". No one could imagine that the data make h_2 certain, so that $p(h_2, eb) << 1$, and the content of h_2—its testability—is apparent from the fact that further characters, those not in the original e, can obviously challenge h_2. In contrast, *ad hoc* h_1, which concerns only the original e, could not conceivably be challenged by further characters. The phylogenetic hypothesis h_2, that is, is definitely not *ad hoc*.

Faith (1992) had previously accused phylogenetic methods of yielding "*ad hoc*" conclusions, but he had used a different argument then and it will be illuminating to compare his positions. According to Faith (1992, p. 267), "*ad hoc*" meant agreement with background knowledge:

> Popper properly defines *ad hoc* explanations as those that rely only on our already-accepted facts (thus, an *ad hoc* hypothesis, in failing to go beyond background knowledge, has low corroboration; see Popper, 1963: 61, 287). It follows that the mpt [most parsimonious tree] would be the *most*, not least, *ad hoc* hypothesis in that it would be most in accord with the background knowledge.

That criticism rested on a fiction, for Popper's (1963, p. 61; underlining added) comment did not concern background knowledge, but rather available evidence:

> One can show that the probability theories of induction imply, inadvertently but necessarily, the unacceptable rule: always use the theory which is the most *ad hoc*, i.e. which transcends the available evidence as little as possible.

Evidently aware of that, Faith (2004) switched to a characterization of *ad hoc* that was more like Popper's (Faith, 2004, p. 8; underlining added):

> I argued earlier [Faith, 2004, p. 5; quoted above] that any goodness-of-fit procedure does nothing more than select an *ad hoc* hypothesis, which (in maximising fit) goes as little beyond the observations (the observed data) as possible.

That argument was not correct either, but Faith's paraphrase of Popper's comment was informative nonetheless. It revealed that Faith knew what he had always denied, that Popper's *available evidence* referred to what others call *observed data*.

Conclusion

It would seem that Faith was not sincere in his repeated insistence that Popper's "evidence" meant fit, that "true corroboration" required fit as "evidence" and so on, but then the same applies to most of his arguments. Considering his pheneticist background, Faith may have been sincere in advocating his verificationist strategy, but even if so he could not possibly have believed that verificationism was Popperian. Nor could he actually have thought that statistical probabilities should not be used in calculating corroboration, for if that had been true it would have barred his own proposed interpretation of statistical probability PTP. He could scarcely have thought that a probability could be negative or that prediction and observation were the same. Nor could he have believed that he was accurately conveying Popper's ideas when he cut Popper off in mid-sentence, or when he avoided mentioning Einstein's theory, or when he edited out the provision for $p(h, b) = 0$. But if the FCT claims about Popper cannot have been meant seriously, they must instead have been advanced merely as pretexts for objecting to phylogenetic methods, in the hope of obstructing the legitimate application of Popper's ideas in phylogenetic systematics.

Faith's arguments were extraordinarily convoluted, but otherwise he was far from alone. If Rieppel actually thought that Popper's ideas were not meant for scientific application, or that Popper did not use statistical probabilities, or that corroboration behaved asymmetrically, this can only have been because he had made no effort to find out—but that was scarcely the impression he gave his readers. If de Queiroz and Poe truly believed that unrealistic background theories could remain unchallenged, then they must have been afflicted by remarkably short memories, but even then they could hardly have thought that they could so argue while objecting to the realism of "NCM", nor even then could they have imagined that their reworded "NCM" really made NCM unrealistic. No if applies to Felsenstein. In pretending that I had not discussed cliques or said how to count homoplasies, and in never mentioning the problems of applying statistical consistency as a criterion in realistic cases, he was simply concealing any information that did not fit his position.

Many, no doubt, would consider such cases depressing, signs of waning integrity among scientists. But there is a silver lining. If these are the strongest criticisms that my view of parsimony has to face, then it has a bright future indeed.

References

Archie, J.W., 1989. A randomization test for phylogenetic information in systematic data. Syst. Zool., 38, 219–252.

Carpenter, J.M., 1992. Random cladistics. Cladistics, 8, 147–153.

Carpenter, J.M., Goloboff, P.A., Farris, J.S., 1998. PTP is meaningless, T-PTP is contradictory: a reply to Trueman. Cladistics, 14, 105–116.

De Laet, J.E., 2005. Parsimony and the problem of inapplicables in sequence data. In: Parsimony, Phylogeny and Genomics. Albert, V.A. (Ed.), Oxford University Press, Oxford, pp. 81–116.

Faith, D.P., 1990. Chance marsupial relationships. Nature, 345, 393–394.

Faith, D.P., 1992. On corroboration: a reply to Carpenter. Cladistics, 8, 265–273.

Faith, D.P., 1999. Error and the growth of experimental knowledge. Syst. Biol., 48, 675–679.

Faith, D.P., 2004. From species to supertrees: Popperian corroboration and some current controversies in systematics. Aust. Syst. Bot., 17, 1–16.

Faith, D.P., 2006. Science and philosophy for molecular systematics: which is the cart and which is the horse? Mol. Phylo. Evol., 38, 553–557.

Faith, D.P., Cranston, P.S., 1991. Could a cladogram this short have arisen by chance alone? On permutation tests for cladistic structure Cladistics 7, 1–28.

Faith, D.P., Cranston, P.S., 1992. Probability, parsimony, and Popper. Syst. Biol., 41, 252–257.

Faith, D.P., Trueman, J.W.H., 1996. When the topology-dependent permutation test T-PTP for a null hypothesis of non-monophyly returns significant support for monophyly, should that be equated with a rejecting a null hypothesis, b rejecting a null hypothesis of 'no structure', c failing to falsify a hypothesis of monophyly, or d none of the above? Syst. Biol., 45, 580–586.

Faith, D.P., Trueman, J.W.H., 2001. Towards an inclusive philosophy for phylogenetic inference. Syst. Biol., 50, 331–350.

Farris, J.S., 1970. Methods for computing Wagner trees. Syst. Zool., 19, 83–92.

Farris, J.S., 1980. The efficient diagnoses of the phylogenetic system. Syst. Zool., 29, 386–401.

Farris, J.S., 1983. The logical basis of phylogenetic analysis. In: Platnick, N.I., Funk, V.A. (Eds.), Advances in Cladistics II. Columbia University Press, New York, pp. 7–36.

Farris, J.S., 1991. Excess homoplasy ratios. Cladistics, 7, 81–91.

Farris, J.S., 1995. Conjectures and refutations. Cladistics, 11, 105–118.

Farris, J.S., 1996. Names and origins. Cladistics, 12, 263–264.

Farris, J.S., 1999. Likelihood and inconsistency. Cladistics 15, 199–204.

Farris, J.S., 2000. Corroboration versus "strongest evidence". Cladistics, 16, 385–393.

Farris, J.S., Kluge, A.G., 1986. Synapomorphy, parsimony, and evidence. Taxon, 35, 298–306.

Farris, J.S., Kluge, A.G., Carpenter, J.M., 2001. Popper and likelihood versus "Popper.*". Syst. Biol., 50, 438–444.

Felsenstein, J., 1973. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. Syst. Zool., 22, 240–249.

Felsenstein, J., 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool., 27, 401–410.

Felsenstein, J., 1979. Alternative methods of phylogenetic inference and their interrelationship. Syst. Zool., 28, 49–62.
Felsenstein, J., 1981. A likelihood approach to character weighting and what it tells us about parsimony and compatibility. Biol. Jour. Linn. Soc., 16, 183–196.
Felsenstein, J., 1983. Methods for inferring phylogenies: a statistical view. In: Felsenstein, J. (Ed.), Numerical Taxonomy. Springer-Verlag, Heidelberg, pp. 315–334.
Felsenstein, J., 1984. The statistical approach to inferring evolutionary trees and what it tells us about parsimony and compatibility. In: Duncan, T., Stuessy, T. (Eds.) Cladistics: Perspectives on the Reconstruction of Evolutionary History. Columbia University Press, New York, pp. 169–191.
Felsenstein, J., 1993. "Phylip." Version 3.5. Computer Software and Documentation. Department of Genetics, University of Washington Seattle.
Felsenstein, J., 2004. Inferring Phylogenies. Sinauer, Sunderland, MA.
Grant, T., Kluge, A.G., 2004. Transformation series as an ideographic character concept. Cladistics, 20, 23–31.
Hennig, W., 1966. Phylogenetic Systematics. University Illinois Press, Urbana, IL.
Hennig, W., 1981. Insect Phylogeny. Wiley, New York.
Hennig, W., 1983. Stammesgeschichte der Chordaten. Verlag Paul Parey, Hamburg.
Kluge, A.G., 1999. The science of phylogenetic systematics: explanation, prediction, and test. Cladistics, 15, 429–436.
Kluge, A.G., 2001. Philosophical conjectures and their refutation. Syst. Biol., 50, 322–330.
Kluge, A.G., 2003. The repugnant and the mature in phylogenetic inference: atemporal similarity and historical identity. Cladistics, 19, 356–368.
Kluge, A.G., Farris, J.S., 1969. Quantitative phyletics and the evolution of anurans. Syst. Zool., 18, 1–32.
Kluge, A.G., Grant, T., 2006. From conviction to anti-superfluity: old and new justifications of parsimony in phylogenetic inference. Cladistics, 22, 276–288.
Laudan, L., Leplin, J., 2002. Empirical equivalence and underdetermination. In: Balashov, Y., Rosenberg, A. (Eds.) Philosophy of Science. Contemporary Readings. Routledge, London, pp. 362–384.
Legendre, P., 1986. Report on nineteenth international taxonomy conference. Syst. Zool., 35, 135–139.
Lindgren, W., 1962. Statistical Theory, Macrnillan, New York.
Popper, K.R., 1959. The Logic of Scientific Discovery, Hutchinson, London.
Popper, K.R., 1963. Conjectures and Refutations: The Growth of Scientific Knowledge, Harper and Rowe, New York.
Popper, K.R., 1972. Objective Knowledge: An Evolutionary Approach. Clarendon Press, Oxford, London.
Popper, K.R., 1983. Realism and the Aim of Science, Routledge, London.
de Queiroz, K., Poe, S., 2001. Philosophy and phylogenetic inference: a comparison of likelihood and parsimony methods in the context of Karl Popper's writings on corroboration. Syst. Biol., 50, 305–321.
de Queiroz, K., Poe, S., 2003. Failed refutations: further comments on parsimony and likelihood methods and their relationship to Popper's degree of corroboration. Syst. Biol., 52, 352–367.
Rieppel, O., 2003. Popper and Systematics. Syst. Biol., 52, 259–271.
Rieppel, O., Rieppel, M., Rieppel, L., 2006. Logic in systematics. J. Zool. Syst. Evol. Res., 44, 186–192.
Siddall, M.E., 2001. Philosophy and phylogenetic inference: a comparison of likelihood and parsimony methods in the context of Karl Popper's writings on corroboration. Cladistics, 17, 395–399.
Sober, E., 1988. Reconstructing the Past: Parsimony, Evolution and Inference. MIT Press, Cambridge, MA.
Steel, M.A., Szekely, L.A., Hendy, M.D., 1994. Reconstructing trees when sequence sites evolve at variable rates. J. Comput. Biol., 1, 153–163.
Tuffley, C., Steel, M., 1997. Links between maximum likelihood and maximum parsimony under a simple model of site substitution. Bull. Math. Biol., 59, 581–607.
Wheeler, W., Aagesen, L., Arango, C.P., Faivovich, J., Grant, T., D'Haese, C., Janies, D., Smith, W.L., Varón, A., Giribet, G., 2006. Dynamic Homology and Phylogenetic Systematics: A Unified Approach Using POY. American Museum of Natural History, New York.
Wiley, E.O., 1975. Karl R. Popper, systematics, and classification: a reply to Walter Bock and other evolutionary systematists. Syst. Zool., 24, 233–243.
Wiley, E.O., 1981. Phylogenetics. The Theory and Practice of Phylogenetic Systematics. John Wiley and Sons, New York.