British Journal for the Philosophy of Science,
55, 2004, 615-643

#182

Cliometric Metatheory III: Peircean Consensus, Verisimilitude and Asymptotic Method

Paul E. Meehl[†]

ABSTRACT

Statistical procedures can be applied to episodes in the history of science in order to weight attributes to predict short-term survival of theories; an asymptotic method is used to show short term survival is a valid proxy for ultimate survival; and a theoretical argument is made that ultimate survival is a valid proxy for objective truth. While realists will appreciate this last step, instrumentalists do not need it to benefit from the actuarial procedures of cliometric metatheory.

1 Introduction

Of the many questions philosophers properly ask in contemplating the aims, methods and progress of science, only a few are of practical interest to the scientist. Suppose, in addition to expounding ideas of intrinsic scholarly value, the philosopher hopes to play the role of *ancilla scientiae* to the working scientist. Far and away the most important question is: 'How should the

[†] This article had been completed by Paul Meehl at the time of his death on 14 February 2003. His wife, Leslie J. Yonce, is grateful to Keith Gunderson (University of Minnesota Center for Philosophy of Science) and Niels G. Waller (Psychology Department, Vanderbilt University) for advice with some final editing details.

merits of competing scientific theories be appraised?' For the present purpose, I am adopting a pragmaticist criterion of theoretical merit, such as Peirce's ([1878/1986], p. 273) famous definition of truth: 'The opinion which is fated to be ultimately agreed to by all who investigate [. . .] and the object represented in this opinion is the real.' As a scientific realist, I take this as a proxy for objective truth-likeness (= verisimilitude), to be theoretically justified below.

As a behavioral scientist, confronted with a puzzling heap of competing theories (most short-lived, and many faddish and of little merit), I have usually been disappointed by philosophers of science writing about this crucial topic. For example, of the numerous properties and relations that theories can have, a typical list in a philosophical paper mentions only three or four, and hardly any writer lists more than half a dozen. Sometimes only parsimony is mentioned, but this favorite is rarely explicated or justified as a criterion of objective truth or long-term survival of a theory, and not everybody accepts it as such, Popper and his disciples valuing it only as a correlate of falsifiability. Further, whatever its merits, there are at least four kinds of parsimony. There are, in fact, at least 18 theory characteristics (Meehl [2002]) that *some* scientists *sometimes* take into account in appraising theories.

Nobody claims that any one theory attribute[1] is a litmus test of verisimilitude or of long-term survival; and, while scientists may have opinions as to which attributes should be weighted more and which less when comparing theories, no one says that any particular attribute will trump each of the others, singly or in various combinations, or all of the others collectively. Philosophers and historians of science rarely show interest in which attributes are in fact given greater weight, nor do they offer a metatheoretical rationale for which *ought* to be preferred. Even Popperians, who claim that falsifiability is the litmus test of a theory having scientific status, do not claim that *degree* of falsifiability as a quantitative matter trumps everything else. Moreover, once a theory has been admitted into the class of being scientific at all, no one, so far as I know, has maintained that no combination of other desirable attributes could properly countervail corroborability.

A naturalized epistemology, which this article presupposes, will presumably claim that if reliance on an attribute is rational it must be because we believe that the attribute correlates with the theory's objective merit, i.e., long-term survival if one is an instrumentalist, objective verisimilitude if one is a scientific realist. It is an *empirical claim* about history of science that, over a given class of scientific theories, or over all empirical theories in all sciences, an attribute is

[1] I shall use 'attribute' to cover both *intratheoretical properties* (e.g., parsimony, mathematization) and *theory relations* (e.g., relations to facts and to other theories). Cliometric metatheory (Faust and Meehl [1992], [2002]; Meehl [1992a], [1992b], [2002]) does not deal with the psychology of the scientist or the sociology of knowledge.

a correlate of survival. We may have plausible—some might claim deductive—metatheoretical arguments for giving an attribute a high weight in our attribute list, but the *content* of the metatheoretical assertion 'attribute A correlates with survival' is empirical. It is intrinsically a statistical claim, as are all alleged correlations; and there is only one way to check statistical claims—namely, to compute statistics. A claim that more parsimonious theories are more likely than less parsimonious theories to survive in the long run, or that theories that generate precise numerical predictions are a better bet than those that do not, or that any of 18 attributes I have listed (Meehl [2002]) are predictive of Peircean survival, are all statistical claims. Whatever one's metaphysics and epistemology may be, I submit that there is no known way to verify or refute statistical claims about empirical relations other than to get the facts and calculate the statistics.[2]

That an empirically based philosophy of science which takes the history of science as its database implies actuarial description of scientific episodes and sophisticated psychometric analysis of those statistics is the *Faust-Meehl Actuarial Thesis* (Faust [1984]; Faust and Meehl [1992], [2002]; Meehl [1983], [1992a], [1992b]). I emphasize that this is a thesis, not merely an expression of a fondness for actuarial method due to Faust and Meehl being clinical psychologists who have worked in the area of prediction. Some philosophers and social scientists seem to have a distaste for this thesis and seem to view it as simply a matter of taste or, at best, a highly subjective personalistic judgment call about strategy. On the contrary, we consider the previous paragraph's short derivation dispositive; I invite the reader to refute it.

2 A plausible proxy for Peircean consensus

Because Peirce's ideal pragmaticist criterion of ultimate consensus is unavailable to us, we must substitute an acceptable proxy in order to implement the actuarial thesis in an empirical research program. Whether it is a good proxy is to be decided by a combination of theoretical (probability) arguments and indirect empirical corroboration. I shall do this without vicious circularity, reminding the reader of Feyerabend's maxim, 'There is nothing wrong with arguing a circle if it is a big enough circle.'[3] I propose as a surrogate for

[2] I do not, of course, deny that one can prove theorems in the formalism of the probability calculus; but the probability calculus as applied to an empirical domain cannot be other than a way of getting from one known or estimated statistical relative frequency to another.

[3] The unavoidability of permissible circularity or appearance thereof is what makes strict positivist phenomenalism an unworkable epistemology, as is, I believe, universally agreed, the last heroic effort being Ayer's *The Foundations of Empirical Knowledge* ([1940]). Incidentally, the maxim is not original with Feyerabend, having been stated by C. I. Lewis in *Mind and the World-Order* ([1929]).

ultimate survival *fifty-year ensconcement.* I will suggest below—bypassing Hume, as we do in naturalized epistemology—an indirect but good test of its adequacy as a proxy. The notion of ensconcement, and its conjugate *discard*, is initially operationalized by a list of social facts about the theory, a list that is here treated qualitatively as a conjunction.[4] The provisional list consists of documentary items such as:

1. In elementary textbooks of the science, the theory is usually the only one presented and is generally spoken of as 'proved', 'established', 'demonstrated', 'no longer a theory but a fact' (a philosophically erroneous expression!), 'generally accepted', 'no longer in doubt', and the like.
2. Advanced technical treatises do the same thing but may present summaries of the compelling evidence or references containing such evidence.
3. Technological works on applications presuppose the theory.
4. The current research literature contains almost no research studies aiming to test the theory, but instead:
   a. Kuhnian 'puzzle-solving' to explain away *prima facie* falsifiers;
   b. technological applications;
   c. efforts to relate the theory to other ensconced theories (reduction upward and downward);
   d. improved articulation of the theory, sharpening of its concepts, more accurate determination of physical constants.
5. Presentations and panel discussions of the theory disappear from scientific meetings except as historical or celebratory.

Treating these five as conjunctive criteria, we assign a date of ensconcement for a theory.[5] We then consider a period of 50 years following the ensconcement date, during which episodes of deviation from the above criteria are of negligible frequency. That fifty-year ensconcement is our proxy for Peircean survival. In the same way, we classify a theory as discarded when it is no longer being mentioned in textbooks except perhaps in referring to its discard (e.g., phlogiston, caloric, miasma, gemmules, phrenology), experiments are not being performed to further refute it, and it has dropped out of the scientific literature except in historical discussions.

Our acceptance of fifty-year ensconcement/discard as a proxy for Peircean truth/falsity might be defended simply by remembering that we subscribe to naturalized epistemology. It would be, if not logically contradictory, at least very odd for someone to say, 'I don't begin with first philosophy, to answer

[4] At a later stage of the cliometric program it might be transformed into a differentially weighted composite; we can't do everything at once!

[5] We do not fret about the exact date, which would violate the maxim of my logical positivist mentor Herbert Feigl, 'Don't cut butter with a razor.' I believe the current variant in physical science is 'Don't cut Spam with a laser.'

skeptics Hume and Co., because I am convinced that cannot be done; I take the view that the philosopher or historian of science should commence by confessing a firm belief that, by and large, most of the theories held with quasi-Quaker unanimity by scientists for long time periods are substantially correct,' but then proceed to reject a two-generation ensconcement as having zero predictive validity for long-term survival or high verisimilitude.

If this somewhat brute-force naturalized epistemology is discomforting, I ask the reader to have patience, to suspend judgment and proceed with provisional acceptance. Philosophers unfamiliar with the practice of bootstraps psychometrics require some indoctrination in a novel way of thinking. In the tradition of psychometricians, we never reject a criterion on the ground that it is fallible. Were we to do that, psychometrics, theoretical or applied, would be impossible. The history of mental testing—of ability, achievement, personality, political and religious attitudes, or whatever—shows that progress can be made starting with criteria of various sorts (e.g., a psychiatrist's diagnosis of a mental patient) that have only moderate reliability and validity.

Of course we know that fifty-year ensconcement is not an iron-clad guarantee that the theory will never be discarded, the classic example being Newtonian mechanics. This one was such a shocker—the theory having for two centuries been the paradigm case of powerful empirical science that everybody wanted to imitate—that it leads some philosophers to say, 'All theories are lies.' For many years, I regularly taught this to graduate students in my Philosophical Psychology seminar to my present embarrassment because it is a mistake. Newton's and other 'Grand Theory' examples have led philosophers in that direction, but I urge that it is currently overdone. Grand theories which, as Feyerabend puts it, 'say something about everything there is,' and which express their content in the precise form of partial differential equations, are especially subject to imperfection. But science contains thousands (yes, I do mean thousands and can prove it) of 'mini-theories' that are not of that grand, all-encompassing sort, and which we can be confident will never be refuted. We know that the liver stores glycogen and secretes bile, which is stored in the gall bladder which empties into the duodenum where the bile emulsifies fat. We know that the genetic code is located in a double helix the backbone of which is a phosphate radical hooked to a five-sugar. We know the sun is not a hot iron ball as thought by Anaximander—a respectable theory given his data—but it is hot gas, mostly hydrogen; nobody is ever going to resuscitate the theory that the sun is an iron ball or Apollo's chariot. Even in the physical sciences where numerical approximations are involved, one can state the theory in a way that frees it from even a faint possibility of being mistaken. We do not have to know exactly the average

distance from the earth to the sun to assert confidently that it is between 90 and 95 million miles.

However, we can supplement reliance on naturalized epistemology by empirical test. Consider a class of theories that achieved fifty-year ensconcement long enough ago so that another half-century has passed since they met that criterion. We now search for any subsequent surprising falsifications (leading to discards) among them. Plotting the cumulative graph of such surprises, we fit a suitable mathematical function to those data and use standard curve-fitting procedures to make a numerical estimate of its asymptote. We do not *assume* that the function has a smoothly approached asymptote, but we know for sure that the percentage of ensconced theories fated to be later falsified has an upper bound at 100%. We then *conjecture* (not 'assume') that, like most biological and social mass phenomena, it is lawful and smooth (not having discontinuities and step functions), so that its numerical 'lid' is approached asymptotically. This conjecture is to be tested by our data. There is a choice of functions to fit and the statistician has ways of doing that.[6] Thus, our belief that fifty-year ensconcement is a good proxy for ultimate survival is not an assumption or an epistemological postulate, but is the result of an empirical fit of a curve to data.

Conjecture: The asymptote of this cliometric function will have a discard probability $p < .05$. This may sound optimistic coming from a psychologist, since the behavioral sciences (they shouldn't be called 'social', because they are not all that) have such a proliferation of discarded theories (Meehl [1978], [1990c]). But one must remember that, correspondingly, very few theories in the less developed sciences, including biology, ever *become* ensconced in the first place, so the high general discard figure does not contradict the small surprise discard (of theories having previously been counted as ensconced) rate that I conjecture. Similarly, we inquire about the surprising resurrection

---

[6] An obvious possibility here would be the so-called simple positive growth function, in which the rate of a process is proportional to how much remains to go—such as the rate of flow of beer out of a barrel, or the growth of a cornstalk. We integrate the equation for this growth process,

$$\frac{dy}{dt} = K(R - y)$$

where y = cumulated amount (e.g., of beer flowed or cornstalk grown), R = original amount or final height, K = a rate constant, obtaining

$$y = R(1 - e^{-Kt}).$$

Fitting this function estimates the asymptote R.

We would consider initially a class of functions $y = f(t)$ such that $f(0) = 0, f'(t) > 0$, and $f''(t) < 0$ everywhere, $\lim f(t) = \mathrm{L}$, that are known to fit data well in the life sciences, adopting the best fitting function, ideally one whose least squares misfitting is statistically insignificant (attributable to random error). Whether t is real time or cumulative number of experiments is decided on the basis of orderliness.

Table 1 Attributes used by scientists in theory appraisal (see Meehl 2002])

$Parsimony_1$: Simplest curve
$Parsimony_2$: Economy of postulates
$Parsimony_3$: Economy of theoretical concepts
$Parsimony_4$: Ockham's Razor (Don't invent a theory to explain a new fact explainable by ensconced theory)
Number of corroborating facts derived
Number of discorroborating facts derived
Qualitative diversity of facts derived
Novelty of facts derived
Numerical precision of derived facts
Reducibility, passive: The theory as reduced
Reducibility, active: The theory as reducer

of slain theories,[7] proceeding in the same way. I predict the asymptote of a cumulative curve of such revivals of discarded theories will be negligible: $p < .01$.

3 Assessing the validity of theory attributes as predictors of theory survival

If these asymptotic conjectures are confirmed, we will possess a *fallible but highly accurate* dichotomous criterion of Peircean survival. We will use it to ascertain empirically the validity of members of the candidate list of attributes known anecdotally to be employed by scientists in theory appraisal. Each one's use, even if occasional or by a scientific minority, warrants its inclusion as a candidate predictor. My list, culled from philosophical writings and my anecdotal material, is given in Table 1. These are the attributes I suggest being tried in the early stages of cliometric study. I select them from a longer list of 18 candidates for four reasons: affirmatively, my impression is that they are more often mentioned by scientists, philosophers and historians of science than the others; I can offer a rough idea of how to quantify each of them; I can offer proofs why each is a plausible correlate of verisimilitude (Meehl [2002]); and none is apparently reducible to a combination of the others. Another seven which do not meet these four conditions, I set aside for evaluation at a later stage in the cliometric research program: initial plausibility, rigor of

[7] Feyerabend ([1970]) attacks Lakatos' ([1970]) notion of degenerating research programs by citing Prout's hypothesis that all the elements are in some sense compounded out of hydrogen, suggested by the near-integer atomic weights in the periodic table. One should be suspicious when a general point is allegedly proved by always mentioning the same example. We should inquire how many such examples there are. Besides, Prout's hypothesis was never discarded in my sense, because all along there were competent scientists who expected that somehow it could be fixed up. Sir William Crookes even predicted that chlorine's anomalous atomic weight of 35.5 would be explained in terms of isotopes (employing the concept before that term had been invented). I am unaware of a single theory in psychology, genetics, or medicine that having been discarded for a half-century was subsequently resurrected in anything like its original form.

theoretical derivations, confidence in the auxiliaries in observational testing, deductive fertility (fruitfulness), technological power, computational ease, and beauty (depth, elegance).

I conjecture that an optimally weighted linear composite of quantified attributes that discriminate the ensconcement/discard criterion will be more accurate with regard to the ultimate fate of a theory than short-term scientific opinion in the theory's history (e.g., at its midlife).[8] This optimistic notion may puzzle philosophers, but it does not surprise a psychometrician; the process of psychometric 'bootstrapsing' has compelling mathematical arguments and a long and varied history.[9]

Taking the fifty-year ensconcement/discard dichotomy as the proxy for ultimate Peircean survival (assuming my scenario's favorable asymptotic statistics), we can subject the eleven candidate predictors to several interesting statistical analyses.

3.1 Linear discriminant function

We can use a linear discriminant function (Fisher [1970], [1971]; Lachenbruch [1975]; Morrison [1990]) to construct a predictive function as a linear composite of the indicators, $y = f(x_1, x_2,\ldots x_{11}) = \lambda_1 x_1+\lambda_2 x_2,\ldots\lambda_{11}x_{11}$, where the λ weights are assigned so as to optimize the separation (discrimination) of the classes of ensconced versus discarded theories. The statistical optimization takes account of the correlation of each predictor with the criterion as well as the $\binom{11}{2} = 55$ pairwise predictor correlations. In beginning thus, we do not assume (what is probably false) that the true function is exactly linear, but experience in the life sciences shows this is a good way to start. Departures from linearity are usually small enough to be neglected, at least in the early stages of investigation (Dawes [1979]; Dawes and Corrigan [1974]), and often remain so. There are two kinds of nonlinearity that one may need to consider at a later stage. In the first, one or more of the variables is subjected to a nonlinear transformation but they are

---

[8] Such a composite function could theoretically exceed the ensconcement criterion itself in forecasting ultimate survival, but we would have no way of predicting this. That ensconcement is a highly valid proxy for Peircean survival does not enable us to say *which* theories will constitute the proxy's misclassifications.

[9] The term 'bootstraps effect' in psychometrics was introduced by Cronbach and Meehl ([1955]) as a provocative phrase to highlight the paradox that one can build mental tests in initial reliance on a highly fallible criterion, which tests may turn out to have higher validity than the initial criterion. Unfortunately, the term 'bootstraps' was subsequently introduced into philosophy (by Glymour [1980]) and into general statistics (by Efron [1979]) with different meanings in each case. Although there is a deep sense in which the three usages are related, I feel entitled to use the word as Cronbach and I first did. A sum of test items each of which was chosen for showing a percent difference between patients diagnosed as schizophrenic or 'normal' can, as a scale, differentiate the groups more accurately than the diagnosing clinician; this is a matter of simple algebra involving the average percent difference and the average inter-item correlations. The classic example of successful psychometric bootstrapsing is the intelligence test.

nevertheless combined additively, e.g., $y = a_1x_1 + a_2x_2^2 + a_3 \log x_3 \ldots + a_{11}x_{11}^{1/3}$. A second type of nonlinearity, likely to be of greater theoretical interest and larger numerical impact, is the existence of interaction terms, as in $y = a_1x_1 + a_2x_2 + a_{1,2}x_1x_2 \ldots + a_{11}x_{11}$. Here the influence of a variable on the prediction is itself dependent upon the value of another variable. For example, in agronomy we may study the effects of fertilizer and irrigation on yield of wheat per soil area, subscripting average yields with and without fertilizer by F and $\overline{F}$ and irrigated or not by I and $\overline{I}$. Then the interaction between fertilizer and irrigation is given by

$$\Delta^2 = (\overline{y}_F - \overline{y}_{\overline{F}})_I - (\overline{y}_F - \overline{y}_{\overline{F}})_{\overline{I}} > 0\,,$$

which means that there is a second-order difference between the first-order differences. If this expression is significantly positive, it tells us that the effect of fertilizer for irrigated plots is greater than it is for non-irrigated plots. Note that removing the parentheses and rearranging terms into new parentheses, we get the same algebraic quantity (i.e., interaction effects are symmetrical). Fertilizer and irrigation *potentiate* each other's influence on yield.

For the continuous case, rather than that typically studied in agronomy or pharmacology where we treat a factor as present or absent, one considers derivatives of a continuous predictor function $y = f(x_1, x_2 \ldots x_{11})$. If a second-order mixed partial derivative

$$\frac{\partial^2 f}{\partial x_i \partial x_j} \neq 0\,,$$

we conclude that the variables x_i and x_j operate *configurally* (Meehl [1954/1996], p. 134).

Not all interaction effects are best expressed as cross product terms, although that is often a good approximation. It would not be advisable to begin a cliometric analysis by setting up a general second-degree equation including all such cross products, because we would be assigning 'optimal' weights to 11 + 55 = 66 terms, a capitalization on sampling error that would make a statistician nervous. For that reason I advocate starting with Fisher's linear discriminant function, saving for a later stage the study of possible interaction effects for pairs of predictors that have survived.

Such discriminant analysis conducted in various scientific domains constitutes an empirical test of the conjecture that each indicator is a correlate of verisimilitude and for that reason functions empirically as a predictor of survival. On that view, the weights would be thought of as reflecting latent correlation with verisimilitude. If, say, parsimony$_2$ got a weight of .30 and numerical precision got a weight of .60, we would infer that the latter is a more sensitive indicator of verisimilitude because it is a better predictor of the survival dichotomy. While the realist is thus motivated, *a consistent instrumentalist requires no such belief or aim*, as the discriminant analysis may be justified from a purely predictive viewpoint. This non-realist orientation is

possible also with respect to all of the cliometric statistics which follow below, despite the realist perspective I adopt in the text. Scientific realism *motivates* the calculations, but they do not *require* it. It would be tiresome to keep repeating that for each of them.

3.2 Factor analysis

We use this procedure to analyze only the correlations between the indicators, omitting consideration of the dichotomous survival criterion. Factor analysis asks what latent mathematical factors can be postulated to underlie the 55 pairwise correlations. The contribution or 'influence' of a factor on an observed variable is called its *factor loading* (the British usage is *saturation*) and may be given a causal interpretation depending on the views and aims of the statistician.[10] For a realist, the important point here is a conjecture that the main *source* (explanation) of correlation between the 11 indicators in almost all of the 55 pairs is that each is a correlate of the latent factor *verisimilitude*. Considering all of the pairwise combinations of the 11 indicators in Table 1, there seems no plausible reason for them to be highly correlated, except for two pairs in which a correlation is to be expected even if they have no validity. Number of corroborating and number of discorroborating facts would be perfectly negatively correlated over theories for which the same number of facts is available; and over theories subject to widely varying numbers of facts, this correlation will be negative in sign, although one cannot predict its size. Parsimony_2 and parsimony_3 will be positively correlated because of constraints imposed on admissible theories. Extremely high or low ratios of number of postulates to the number of concepts would result in postulate sets that are redundant, inconsistent, or deductively unfruitful; here again, we anticipate that there will be a relationship, but we cannot say how large it will be. I suggest that for the other 53 indicator pairs, the only plausible reason for expecting substantial correlations is *a priori* arguments (given in Meehl [2002]) as to why each should be a truth correlate.

Optimistic predictions for the cliometric results are that the factor analysis will reveal one big factor that accounts for almost all of the variance of the system, and that the profile of the 11 factor loadings on that first factor will closely match the profile of the criterion discriminant weights (found by linear discriminant function). Why else should we expect a high profile similarity, when the survival criterion dichotomy was deleted from the factor analytic data set? Finding such a high profile matching tends to corroborate our metatheoretical conjecture about verisimilitude.

[10] For an introduction to factor analysis, see Gorsuch ([1983]) and other references given with comments in Meehl ([1999b], note 11, p. 289).

3.3 Taxometric analysis

Intuitively, one thinks of true theories (in the ideal case or in the approximated case of very high verisimilitude) as somehow qualitatively different from the whole class of false theories whose members, however, may contain elements of truth. Consider the true theory of a restricted domain. Set aside earlier versions of it, which in the course of a Lakatosian research program will have undergone progressive improvements in verisimilitude. The false theories, despite sometimes (not usually) containing elements that are similar to the true theory in some pre-analytic sense,[11] are plainly erroneous. In the extreme case one does not ask about the numerical parameters or the mathematical functions in the postulates because the postulates cannot be set into any sort of correspondence with those of the true theory. They are structurally such that by implicit definition (e.g., Ramsey sentence) they do not even, so to speak, postulate existence of the *right sorts* of entities and processes (e.g., there is no caloric, rather there are molecules in motion). However, short of that extreme of zero verisimilitude, for theories that occupied scientific research attention for an appreciable period of time there will be several theories that were once in competition with the true theory and that had some ontological merit, whatever may have been their epistemological status at a given stage in terms of the 11 predictors.

We then face the problem of how to detect the existence of a difference in kind over and above differences in degree. The life sciences have been faced with this problem for a long time: is a collection of entities, whose members differ qualitatively and quantitatively in observable respects, composed of individuals having merely different coordinates in the hyperspace of quantitative dimensions, and without a discernible clumping in the probability density reflecting the existence of a *latent class* (type, species, category, natural kind, *taxon*)? Or is the pattern of relationships of the ascertainable dimensions indicative of a latent taxon, discriminated by the manifest indicators? Suppose the answers to those questions are negative and affirmative respectively; that is, the statistical relations of the manifest indicators corroborate the conjecture of latent taxonicity. We then ask what proportion P (called the *base rate*) of the population belongs to the taxon and what proportion ($Q = 1 - P$) to the complement class. A half-dozen different statistical approaches exist for doing this (see, e.g., Grove and Meehl [1993]; Meehl [1973], [1995], [1999a]; Meehl and Golden [1982]; Meehl and Yonce [1994], [1996]; Waller and Meehl [1998]).

Over a good-sized class of factual subdomains, in each of which there exist multiple theories, only one theory per subdomain can be correct. Identifying subdomains in a science (e.g., human physiology), we can count the number of

[11] This is the kind of unavoidably rough notion that a cliometric research program gets along with in the meantime, with the intention of precisifying at a late stage.

proposed theories in each of those subdomains.[12] Dividing the number of ensconced theories by the number proposed should, on our verisimilitude interpretation, yield a number close to the taxometrically estimated base rate.

In the course of testing a taxonic conjecture, one obtains by indirection certain latent numerical values, along with results from a battery of *consistency tests* which indicate whether one should be confident of one's answer to the taxonic question. These consistency tests also indicate whether confidence may be placed in the several numerical values inferred about the latent situation.[13] An appropriate measure of the amount by which each of the 11 indicators separates the taxon and complement classes should yield a 'validity profile' that closely matches the profiles of the first-factor loadings and the discriminant function weights.

4 Verisimilitude Index

The concept of verisimilitude (truth-likeness, nearness to the truth) is in current disfavor among some philosophers of science, but in my opinion this attitude is mistaken. The objection seems to be that Popper and Co. have not succeeded in constructing a rigorous explication of the concept, even to their own satisfaction, let alone one commanding Quaker consensus. Let me say briefly why I disagree with the rejective attitude. If metatheory is taken to be the empirical scientific theory of scientific theorizing and we operate in the broad framework of naturalized epistemology, then our rational reconstruction of scientific thinking should map the sociopsychological facts of scientific behavior. In my work with scientists in diverse areas,[14] I find that they all take the verisimilitude concept for granted, whether or not they use the word or have ever heard of Sir Karl Popper. In dealing with theory, scientists routinely—not just at moments of great theoretical crisis or Kuhnian revolutions—speak of improving a theory, say that both of two theories are of course imperfect but one is better than the other, ask whether a theory has so little truth that it would be better to go back to the drawing board and start afresh, hope in the long run to improve an ensconced theory that is close to

[12] This latter would be a very close approximation to the total number of entic theories (those that have been or ever will be invented) except for ensconced theories later deposed. No further theories are likely to be invented, since scientists do not practice the 'maximum proliferation' policy advocated by Feyerabend.

[13] These values include base rate, separation of latent means, variances, and location of the HITMAX cut on each indicator (i.e., the cut that minimizes errors in classifying the individual elements into taxon or complement membership).

[14] In addition to a half-dozen fields within psychology, I have done interdisciplinary research and seminar instruction with geneticists, neurologists, psychiatrists, sociologists, political scientists and statisticians. Hence, I have experience with how biological and social scientists think about the verisimilitude of theories. Reading the history of physics, chemistry and geology reveals no qualitative differences.

perfect although not literally true, and maintain that if two theories differ only in the parameters in their postulated functional relations, one is truer than the other. They consider that last sort of difference to be not as great as when theoretical variables are wrongly linked as to what causes what. Still worse are theories that postulate entities and processes 'of the wrong kind' (caloric, phlogiston, miasma, cerebral protuberances representing personality traits, dammed-up libido flowing backward).

One need not pursue philosophy of science to see the unavoidability of closeness to the truth as a meta-concept of rational discourse. In social and business life, in courts of law, in appraising the work of journalists, historians and biographers, we constantly recognize that some narratives are literally accurate, others contain few errors, others numerous minor errors, while other narratives contain major errors and some accounts are wholly fictional. We recognize that scientific theories are not single atomic sentences but inter-related systems of sentences, that these sentences often express quantitative functional relations between states of entities, and that those functions contain numerical constants (parameters). Obviously some numerical values can be more accurate than others. It would be absurd to deny that some theories are better than others. Popper's student D. W. Miller ([1975]) offers an alleged rigorous proof that if two theories are false it is meaningless to rank them as to verisimilitude. If I cannot see what is wrong with his proof, I will nevertheless say that he cannot be talking about the way science actually works, so that whatever concept of verisimilitude he is relying on in his proof must be an unsatisfactory one.[15]

The overarching metaprinciple relied on seems to be that no metaconcept should be permitted in metatheoretical discourse unless it has been rigorously explicated in mathematics and symbolic logic to everybody's satisfaction. No cliometric research is needed to realize that philosophers and historians of science do not normally operate under any such perfectionistic principle. There are many widely-used terms that have not been rigorously explicated to everybody's satisfaction, or even in most cases to the complete satisfaction of the explicator; examples include: truth, induction, proof, disproof, confirmation, falsification, convention, implicit definition, causal nexus, mental events, intentionality, reference, reduction, probability, dispositions, possible

[15] Naturalized epistemology considers it inappropriate to criticize scientific practice in terms of 'first philosophy', and I agree with that principle. I do not, however, hold that no accepted practice of a science may ever be properly criticized from an epistemological standpoint. The widespread abuse of null hypothesis significance testing under the delusion that it can provide severe tests of substantive causal theories is an example (Harlow, Mulaik and Steiger [1997]; Morrison and Henkel [1970]); although some components of that criticism involve disagreements about statistics, the fundamental criticism is philosophical in nature. A careful formulation of a philosophers' defense principle might read something like: 'If a practice, concept or procedure of mature science appears to play an essential role in its success but seems to lack epistemological justification, one should consider it a puzzle demanding solution.'

worlds, bridge laws, operational definitions, psychological causality, determinism, analytic, synthetic, contingent, necessary, synonymy, natural kind, intuition, protocol, observable, best explanation, use novelty, convergence, bootstrapsing, background knowledge, sense data. Likewise, we cannot get along without the concept of verisimilitude, even if we have to rely in the *interim* on such a crude, pre-analytic explication as 'closeness to the truth'.

Having concluded that the metaconcept of verisimilitude is indispensable despite not having a symbolic logic explicans, our pre-analytic intuitive grasp of the explanandum suggests several desiderata for an adequate explication, and hence for the content and (rough) quantitative attributes of an acceptable index. Although a simple count of number of true postulates will not suffice, that idea must nevertheless be somehow captured, as shown by considering extreme cases. Considering two twelve-postulate theories, if T_1 contains eleven true postulates and one false, and T_2 consists of one true and eleven false postulates, an index which fails to score T_1 as higher in verisimilitude would be grossly defective.

Second, when a postulate asserts a functional relation between quantitative values of theoretical entities, $\theta_1 = f(\theta_2)$, there are degrees of specificity in characterizing the function. A weak specification would say that $f'(\theta_2) > 0$ everywhere. If we say further that $f''(\theta_2) < 0$ everywhere, we claim that the function is monotone increasing decelerated. A further specification, identifying a subclass of such functions, would be to state the function form (e.g., $\theta_1 = a + b\log\,\theta_2$ or $\theta_1 = c + d\sqrt{\theta_2}$). Finally, we may state the numerical values of the parameters, e.g., saying $\theta_1 = .73 + .15\,\log\,\theta_2$. Should a postulate be penalized for being more specific but erroneous, receiving a lower score than it would have received had it been less specific? I do not know. That is the sort of question to be answered at a late stage of the cliometric program, presumably by combining empirical and theoretical considerations.

Third, Lakatos' (unexplicated) distinction between core and peripheral components of a theory is a necessary feature, improving on a simple tally of true and false postulates. Like verisimilitude itself, 'coreness' or 'centrality' is a quantitative concept (surely a matter of degree, although Lakatos does not say) and is indispensable, despite its vagueness. In examining a scientific theory, from the most advanced physical science to the least developed social sciences, one sees immediately that some notions are core and others peripheral. For example, if a psychologist claimed to be a 'neo-Freudian', the 'neo' might denote reservations about the importance of penis envy in female neurosis. It might even mean a rejection of that notion. One could cavil about entitlement to the prime word 'Freudian' in such a case, but it would be an arguable matter, somewhere in the semantic gray region. But if a psychologist said 'I am a neo-Freudian,' and meant 'I amend Freudian theory by not accepting the concept of unconscious mental processes,' that would be nonsense. Similarly, if a psychologist said, 'I am a neo-Skinnerian,' that

might legitimately mean adding a special postulate to handle the recalcitrant phenomenon of latent learning (MacCorquodale and Meehl [1954]), whether or not one thought Skinner would approve of it. But if, by 'neo-Skinnerian', one meant disbelief in the concept of reinforcement, that would be an abuse of language.

Accepting these examples as compelling, can we begin to spell out the core/peripheral dimension? Not well, but a little. Contemplating the internal relations among the postulates and the fact domain, one can sometimes see that a postulate is going to be highly *pervasive* in derivations of operational formulas (Meehl [1990a], [1990b]). Roughly, suppose numerous postulates relate different theoretical variables (θ_i, θ_j, θ_k . . .) directly, or through a derivation chain indirectly, to variables θ_1 and θ_2 and $P_{1,2}$ relates θ_1 and θ_2, then that postulate is highly pervasive. For example, in Clark Hull's ([1943]) old theory of mammalian learning, there occur various postulates concerning the concepts conditioned inhibition, stimulus generalization, and other special variables appearing in experimental contexts. All of these contexts involve reaction potential ${}_sE_r$, which depends on habit strength ${}_sH_r$, which in turn depends upon number of reinforcements. Consequently, the basic acquisition postulate concerning the growth of habit strength is highly pervasive.[16] One learns in elementary mechanics about subdomain laws concerning friction, elasticity, capillarity, specific gravity, and the like; but f = ma and the conservation of energy are pervasive postulates. These examples suggest an *a posteriori* approach to pervasiveness, in which the cliometrician would randomly sample from the experimental literature to estimate the proportion of derivation chains to operational formulas in which each postulate played an essential role.

Not being a logician, I have not attempted to construct a symbolic logic explication of verisimilitude; and I am skeptical about the fruitfulness of that way of going about it *at this time*.[17] This is not to say that an acceptable interim explication may violate the laws of logic, but rather that symbolic logic concepts do not provide adequate tools for the task. I think we should collect oral and written quotations from working scientists who speak of better and worse theories and, by content analysis of such discourse, try to tease out the qualitative and quantitative features of theories to which they are pointing. I believe that a list of such informal respects in which a theory can match or fail to match the true theory is a better way to go about it. A provisional index of verisimilitude—admittedly crude and lacking thorough logician-type articulation—could serve as a useful measure in cliometrics.

[16] In this case, it is literally ubiquitous, because any experiment studying any of the other postulates has to begin by equating habit strengths.

[17] I know, however, that the Finnish logicians have not given up (e.g., Niiniluoto [1998]) and may succeed in doing it yet; this is not to say that I consider the symbolic logic approach as the best way to go at it.

Table 2 Progressively stronger specifications in comparing two theories (similitude) (Meehl [1990b]; adapted from an earlier version published in Meehl [1990a])

Theory Specifications	
I.	Type of entity postulated (substance, structure, event, state, disposition, field)
II.	Compositional, developmental or efficient-causal connections between the entities in I
III.	Signs of first derivatives of functional dynamic laws in II
IV.	Signs of second derivatives of functional dynamic laws in II
V.	Ordering relationships among the derivatives in II
VI.	Signs of mixed second-order partial derivatives (Fisher 'interactions') in II
VII.	Function forms (e.g., linear? logarithmic? exponential?) in II
VIII.	Trans-situationality of parameters in VII
IX.	Quantitative relations among parameters in VII
X.	Numerical values of parameters in VII

Roughly, such an index would ask of each postulate in a theory whether it is concerned with the right kinds of entities (caloric fluid versus moving particles, conditioned reflexes versus dammed-up libido), and whether the causal and compositional relationships between the theoretical entities are correct. For example, we do not want every statistical connection to be written as a causal arrow; the latter gives rise to the well-known problems of path analysis in the life sciences (Meehl and Waller [2002]; Waller and Meehl [2002]). We would ask whether a function $\theta_2 = f(\theta_1)$ has the proper algebraic sign for the first derivative, for the second derivative, and so on. A correct theory has all these various kinds of relationships, including the numerical values of the parameters that appear in the various functional relationships. For a large class of theories that could be stated in terms of postulates of this kind, Table 2 lists levels of specificity at which the postulates of a theory can be correct, incorrect, or unstated. I do not of course offer this list as appropriate for all kinds of scientific theories; for some of them it would be inapplicable.

How should we deal with the unavoidable and, for cliometric purposes, important metaconcept of verisimilitude, given its present imperfect state? If we are consistent adherents of naturalized epistemology, viewing metatheory as basically like other empirical scientific theories, we will view its fuller explication as an aim of our metatheoretical research program. Meanwhile, we treat it as scientists customarily treat theoretical constructs, beginning with our pre-analytic, intuitive, commonsensical notion that some theories have a greater truth-likeness than others. We conjecture that there is some sort of truth-likeness *dimension* running through classes of theories and we attempt to get a handle on this dimension by the actuarial and psychometric procedures described above. To the extent that these different epistemic paths to verisimilitude cohere in the manner predicted, we have supplemented our pre-analytic notion by a kind of implicit definition of the construct which,

while much 'looser' than the implicit definitions of a formal science such as geometry, is basically similar to the implicit definitions that occur in first-order empirical scientific theories.

With this approach, we think about metatheory as scientists routinely think about first-order theories in various fact domains, reasoning thus: if there is a dimension of verisimilitude on which theories differ, it manifests itself in the factor analysis of our eleven predictors, in the linear discriminant function for predicting fifty-year ensconcement/discard, and in the taxometric analysis of the indicator patterns. We are encouraged in this interpretation by antecedent theoretical proofs showing *why* each indicator, under very general conditions (bypassing Hume), should rationally be expected to be a truth correlate (Meehl [2002]). All this is nothing new to the scientist, and if a philosopher finds it all wrong-headed, I would suspect an incomplete commitment to naturalized epistemology.

The openness of open concepts (Pap [1958]) has three aspects which are distinguishable but related (Meehl [1977]). Taking a concept to be defined implicitly by its role in the postulated network where the nodes are the theoretical entities and the strands are their postulated relations (causal and compositional), the openness of a concept arises because (a) the net is incomplete, (b) some of the strands are stochastic rather than nomological, and (c) we intend to fill in the nodes by an explicit definition in terms of their composition ('inner nature'). For verisimilitude, this third aspect suggests an explicit definition of truth-likeness in which we compare the 'innards' of a theory with the perfectly correct theory, T_T. We would like to construct an index of objective verisimilitude with T_T as the ideal criterion and then see how the index works cliometrically.

We apply the verisimilitude index to theories in various stages of their evolution (especially at midlife), treating a fifty-year ensconced theory as proxy for truth. Spelled out, ensconcement is a good proxy for ultimate Peircean survival, which is a good proxy for objective truth, as will be shown below. I have not invented an index that would avoid all the well-known difficulties, but I do have some suggestions which begin with reflection on how working scientists talk about verisimilitude. We would like to have a numerical index that behaves properly and, if possible, is standardized so that algebraically it lies between 0 and 1. We want it to be a pure number that will not depend upon the physical dimensions in different sciences. While it is neither a probability nor a correlation, it is a familiar scale, like the numerous pure-number indexes of closeness of relation between variables in the life sciences.

The general theory (epistemology-ontology-mathematics) of index numbers is complicated, controversial and beyond the scope of this article. I am using 'index number' more broadly than its original use in economics; I use it here to mean a numerical composite of two or more quantities to yield a single number that purports to numerify, however crudely and vaguely, a

property, process, or class of entities that we conceive to exist in different amounts, but that—for whatever reason—we cannot assess by direct measurement. Unfortunately for conceptual clarity, one reason for this inability to get a direct measurement may be that no such objective quantity exists. Even in that extreme case, we may have good reasons for conveying the numerical information as a single number that in some sense 'summarizes' the component variables. A common rationale is knowledge or belief that the variables (a) act on the same output, (b) depend causally on the same causal input, (c) have similar observable properties (semantic overlap), and (d) are strongly pairwise correlated. A human personality trait, made up of phenomenologically related and statistically correlated facets (aspects, manifestations, 'atomic dispositions'), is an example. We may have conjectures about its physical nature (e.g., what exists in degree is literally a count of qualitatively homogenous entities—molecules, polygenes for the heritable component of intelligence, silver dollars), or we may rely on the mathematical relation between statistical factors identified in a factor analysis. In the life sciences, it may happen that a causal factor identified statistically by various input-output relations may be qualitatively heterogeneous. For example, communication is a factor in warfare and the efficacy of transmitting information would doubtless emerge as a statistical factor in an analysis of battles. The quantitative determiner of communication adequacy is measurable in information theory terms, but the physical mode of transmission may include such qualitatively diverse processes as carrier pigeon, heliograph, field telephone or motorcyclist (Meehl [1993]).

In constructing a verisimilitude index to quantify the resemblance of a theory T_i to ensconced theory T_{50} (eventually to be able to compare T_i with competitor T_j), we do not commit ourselves to a theory of verisimilitude that makes claims about homogeneity of elements or aspects. Nor do we begin with an index that assigns *a priori* different weights to the levels of specification in Table 2. (A statistical mini-study of top caliber research psychologists revealed zero average agreement among them as to the importance attached to those ten levels; Meehl [1992a].) Our task is to *concoct* (the best-flavored word here!) an index that will numerify our pre-analytic intuitions concerning the goodness or badness of theories in order to get the criteria-matching aspect going. We have no illusions about factorial purity or psychometric optimality in the eyes of Omniscient Jones, and our intention is to modify the provisional verisimilitude index as the cliometric program develops empirically. We do not require ourselves to list evaluative criteria in advance of the program's development, except for the ubiquitous guideline of scientific method as to choice of a metric, namely, *to increase orderliness*.

The basis of our index is the hierarchy of specificity levels in Table 2. Looking at the postulates of candidate theories T_i and T_j at their respective mid-lives and the postulates of criterion theory T_{50} (a theory that ultimately

survived in its specified fact domain), we want an index to express how similar they are. I shall illustrate with a candidate index that I call VINDEX (Meehl [1992a]).

Obviously a theory gets credit for matching a postulate with T_{50}. It gets no credit when it fails to match one, and it loses credit for containing a postulate not in T_{50}. If theory T has some postulates that match T_{T}, but other postulates are missing, then T must have VINDEX < 1; if, in addition, it has false postulates, we want its VINDEX to be still less. If a postulate in T matches one in T_{T} as to the kind of entity and relation to others in the net, but has less specificity as measured against the ten levels listed in Table 2, that should be reflected in a still lower value of VINDEX. On the other hand, it would seem intuitively odd to penalize a theory because its false postulates are more specific than if they were less so. Moreover, if we adopt a metric 'punishing' T_i for false postulates, that would lead to negative values. Let us stipulate the range of VINDEX from 0 to 1. To assure that outcome, we will not subtract false postulate tallies in T_i or T_j in the numerator of our VINDEX fraction; rather, we will include that tally in the denominator. Thus, if n_1 = number of unmatched postulates in T_i, n_2 = number of postulates in T_{50} missing from T_i, and $n_{1,2}$ = number of postulates shared, then our first approximation based on a simple postulate matching tally would be $\dfrac{n_{1,2}}{n_1 + n_{1,2} + n_2}$, similar to a formula in psychometrics for the correlation of two mental tests in terms of overlapping elements.

Using the ten specificity levels in Table 2, we can elaborate our index by counting for each shared postulate the degree of quantitative specification, which gives rise to a range of numbers for each shared postulate from 1 to 10. We refrain from punishing false postulates on the basis of their specificity; that is, a false postulate treating a function as logarithmic is not counted against the theory any more than a false postulate, stating less specifically, that it is monotone increasing decelerated. Generalizing for future refinement, we rewrite the above formula as

$$\frac{n_{1,2}}{\mathrm{A}n_1 + \mathrm{B}n_{1,2} + \mathrm{C}n_2}$$

The important point is that, absent derivation of a unidimensional latent variable (which we have no good reason to believe exists), we are concocting an index comparable to indexes routinely employed in the life sciences. They have a large element of conventionality but are not wholly arbitrary, because they do attempt to quantify (however roughly) an intuitive dimension or kind of difference. The practicing economist cannot get along without the Consumer Price Index, the sociologist and social psychologist need the Index of Social Class, the World Health Organization indexes countries' medical care, the United Nations uses some 44 indicators to index countries' quality

of life, rehabilitation medicine indexes degrees of impairment due to illness, and the like. Unless all such crude indexes are despised on principle, the historian or philosopher of science should not reject a VINDEX composite on grounds of roughness or imperfect mathematical justification.

We might here take a hint from the industrial psychologist. When a job analysis (say, for being a fighter pilot or a clerk in an old-fashioned multi-purpose drugstore) finds the job to involve a heterogeneous list of abilities, knowledge, skills and personality traits, it may be impossible to assign weights to the components that will be agreed upon by hirers or management. In such cases, one approach is to devise a 'most predictable criterion' (Hotelling [1936]) in which we assign weights to the components of the predictand such that an optimal weighting of the predictors (such as a battery of mental tests) predicts this composite criterion better than any differently-weighted composite criterion could be optimally predicted by the tests. That may or may not seem sensible in the military or industrial context (where the statistical procedure is known as *canonical correlation*), but it has a somewhat greater intuitive plausibility here. We do not know whether there exists a nonarbitrary way of combining the truth, falsity and quantitative specificity of scientific theories into a single measure of truth-likeness. But if we have identified a latent quantitative dimension mathematically, as underlying a pattern of correlations, then an index of theoretical closeness that has a good correlation with that implicitly defined dimension would be acceptable until something better comes along.

5 Satisfying both instrumentalists and realists

Assuming the various correlations come out as I optimistically predict, where do things stand with respect to realism and instrumentalism? What we have is a complicated mixture of explanation and justification, as is usual in first-level empirical science. At the fact level, our asymptotic result validating fifty-year ensconcement/discard as a proxy for ultimate survival/discard is not a rash epistemological assumption but an empirical finding. An instrumentalist, accepting Peircean ultimate survival as the definition of truth, relies on the asymptotic result; thus the modest epistemic aim is guaranteed directly. In that sense, the cliometric program fits nicely into an instrumentalist frame. What follows next from a realist perspective does not amend, retract, qualify or contradict a purely instrumentalist interpretation of the cliometric statistics; a consistent (non-dogmatic) instrumentalist would see it as simply a pointless addendum.

How about the realist? Here the aim is verisimilitude, for which ultimate survival is a proxy. The realist could defend this in two ways: within the framework of naturalized epistemology, it would be incoherent to say that theories that ultimately survive are no more likely to be substantially correct than those that do not, so that for one who operates within the framework of naturalized epistemology that *might* be sufficient. Accepting the practice of

psychometric bootstrapsing (Cronbach and Meehl [1955]) and its mathematical rationale, we argue thus: 'Science is the securest sort of knowledge we have. No other doxastic enterprise is in the running with it. Trusting that ultimately surviving theories are almost always essentially correct, and employing a highly valid proxy for this outcome, we determine the validity of 11 theory attributes statistically. We find that an index of competing theories' *content* closeness to the ensconced theory correlates highly with the survival composite. The "best explanation" of these results is that VINDEX and the latent factor are measures of verisimilitude.'

Not wishing to rely wholly on naturalized epistemology (even if one subscribes to it), we may seek an *a priori* proof that false theories will be detected. We want to get a reasonable estimate of the proportion of false theories that will be detected as false in the very long run. In Peirce's ideal pragmaticism, these theories are not 'fated to be ultimately agreed on by all who investigate'. If false theories are quasi-certain to be detected in the long run, those theories that do survive are quasi-certainly true. If the asymptotic method has shown fifty-year ensconcement to be a good proxy for survival, that (available) criterion is a near guarantee of objective truth. This may sound unattainably grandiose to a philosopher, but I shall have a go at it.

There are two classes of false theories to consider. First, false theories are obtainable by substituting an incorrect mathematical function for the true one. The corrupted theory T'_{T} is, so to speak, topologically like T_{T} in that the nodes and strands of its nomological network are the same, and the interpretive text, if any, characterizing the entities and processes is the same. Corrupted theory T'_{T} postulates the same kinds of entities and the same causal and compositional relations between them as does T_{T}. But perhaps where T_{T} has a logarithmic function connecting a state, property or event in theoretical entity θ_1 with a state, property or event in θ_2, T'_{T} has instead a square-root function. All of the operational formulas in whose derivation chain this theoretical postulate plays an essential role will be incorrect. If our measuring instruments are sufficiently precise, T'_{T} will be asymptotically detected; that is, as the community of scientists persists in sampling the population of operational formulas derivable from T'_{T}, the probability approaches zero that all the incorrect ones will remain untested. If scientists continue to sample the various derived operational formulas of the fact domain until the sun burns out, the probability of falsifying $T'_{\mathrm{T}} \rightarrow 1$. Hence, this class of incorrect theories are all quasi-certain to be falsified in the long run.

The other kind of false theory is not network-isomorphic with T_{T}, not merely a matter of selecting a wrong mathematical function while having the same causal and compositional structure. Rather, it postulates different kinds of entities or draws the wrong causal connections between the nodes in the net. Bypassing whether logicians are correct in saying there is an infinite number of theories adequate to derive the complete domain of observational

formulas, I limit myself to a class I take to be finite, however large, of false-but-adequate theories which are undetectable by the asymptotic method, no matter how many experiments we perform. This class of qualitatively incorrect but factually adequate theories may be either the admissible class (meeting some constraints on structure, content or compatibility with background knowledge) or the accessible class (having a non-negligible probability of conception by some scientist; see Meehl [2002]), so long as it is finite. Let the number of mathematical functions accessible be k. If a theory of n_i postulates is accessible, then any of the k_i conceivable mathematical functions is accessible for each postulate. The number of false theories is $\overset{N}{\Sigma}\, kn_i$. The number detectable because corrupted is $\overset{N}{\Sigma}\,(kn_i - 1) = \overset{N}{\Sigma}\, kn_i - \mathrm{N}$. Hence, the detectable proportion is

$$\frac{\overset{N}{\Sigma} kn_i - N}{\Sigma kn_i} = \frac{N\overline{kn_i} - N}{N\overline{kn_i}}$$

$$= 1 - \frac{1}{\overline{kn_i}}$$

Suppose there were only $k = 10$ accessible functions, and the average number of postulates per theory were $\overline{n}_i = 15$. Then the proportion of detectable false theories would be $1 - 1/150 > .99$. This is of course a gross underestimate, because we are considering only one postulate corruption per theory, whereas one might corrupt 1, 2, 3, …, n postulates by selecting various subsets. (This large number of possible corruptions is only faintly attenuated by the few multiple corruptions which exactly countervail one another.) Taking this value of detection probability together with that for T'_{T} corruptions, we conclude that the asymptotic detection probability over the whole class of false theories is very close to 1, a quasi-certainty by Buffon's famous definition ($p = .9999$).

That many false theories have been concocted and, for varying lengths of time, believed by scientists, has been used to draw pessimistic conclusions about scientific progress and, derivatively, unwarranted metaphysical inferences against scientific realism. I think this unjustified pessimism arises from the unfortunate tendency of philosophers to focus on the shocking overthrow

of a few Grand Theories, the most scandalous example being Newton.[18] The quasi-certainty that Peirce's proxy will 'work' cannot assure us that science will reach the truth in the very long run, but only that it will not persist in theoretical error. Obviously no armchair showing is possible that objective truth is always accessible. It may be that some true theories are inaccessible due to the limitations of the human mind. I can think of no way to get partial reassurance on this score, except by cliometric research.

Pending such research estimating theoretical success rates in various scientific domains, let me take a reassuring example from a discipline intermediate in scientific development and rigor between sciences like astronomy, chemistry, and physics, and, say, psychoanalytic theory or cultural anthropology, which some would say have negligible claims to being sciences. Consider the life sciences, specifically theories about the human body. This fact domain has been assiduously explored because of its anthropocentric theoretical interest and its practical importance, resulting in a large number of mini-theories. Each organ in the human body is a candidate for theoretical explanation concerning its structure (anatomy, histology and microanatomy), the physical chemistry of its function, and its 'teleological' role in the economy of the organism. If we count all striped muscles as one organ, all smooth muscle as another, all arteries as another, and so on, there are by my count approximately 100 different organs in the human body. Except for those considered vestigial (e.g., the vermiform appendix), a satisfactory theory exists for all of them. That does not mean that no further research is going on, but that the research is largely of a clean-up, puzzle-solving, and numerically precisifying kind. We do not have serious doubts that these mini-theories of structure, process and role will stand the test of time. Despite the example of Newton's theory, no anatomist or physiologist would hesitate to wager that a hundred or a thousand years from now scientists will still agree that the liver stores glycogen and secretes bile. We can be equally certain that the thickening of the eyeball's lens takes place in a clever indirect manner in which the ciliary muscle *reduces* lateral pulling by the suspensory ligament, so that the lens thickens from its own inner forces.

Entic theories (the class of theories that are in fact ever concocted; see Meehl [2002]) are probably not a random sample from accessible theories, the selection of false ones being biased in favor of non-corrupted (and hence undetected) adequate variants. But it seems unlikely that we are so clever as to concoct this preferred subclass in a ratio of 10:1 or 100:1, being supernaturally inspired by Ahriman (or Ormazd, for instrumentalists). Even a

[18] It is a mistake to discuss realism as if the only science were physics and—worse still—as if the only branch of physics were quantum theory, which has been conceptually unsatisfactory in its various phases for four generations of physicists. Until I am reasonably clear as to what theoretical statements mean, or even as to whether they purport to denote anything at all, I cannot interest myself in whether they match reality.

100:1 bias would only damage our detection probability from .9999 to .99, a perfectly acceptable reduction.

6 Recapitulation

The novel and deviant character of cliometric metatheory conduces to misunderstandings by philosophers, thus a recapitulation of the above reasoning seems appropriate. Omitting the *a priori* justificatory arguments and assuming the optimistic statistical results imagined above, the cliometric procedures, findings and inferences therefrom proceed sequentially as follows:

1. In the biological ecosystem mind-and-society-in-the-world, scientists invent theories to explain their observations, and the community of scientists appraises these theories as to their merits. In the long run, most theories are rejected for various reasons, the chief and ultimately determinative reason being the inability to explain all the facts.
2. We conjecture that the system is orderly rather than chaotic, although the 'laws' of the system are statistical rather than nomological. This conjecture is empirically testable by cliometric statistics.
3. Appraisal of theories is not based upon strict rules concerning sharply defined findings, but by a loose composite of theory characteristics (properties and relations) to which different scientists give different evidentiary weights.
4. I have listed eleven such characteristics found in scientific articles, works on history and philosophy of science, observations of other scientists, and introspections, and I have elsewhere (Meehl [2002]) given plausibility arguments for their statistical correlation with theory truth.
5. The ultimate survival of a theory (as in Peirce's pragmaticist definition of truth) should be predictable to some extent at the theory's mid-life by a suitable mathematical combination of its quantitative scores on these characteristics.
6. That ultimate survival not being known to us, we adopt as a proxy a dichotomy defined as fifty-year ensconcement/discard.
7. Ensconcement and discard are each defined by a conjunction of features showing the scientific community's treatment of the theory as 'proved' versus 'discarded' persisting for a half century.
8. Because ensconced theories are sometimes discarded in the long run, we plot the curve of such surprising reversals and take its asymptote as the error rate for ensconcement as a quasi-criterion of survival.
9. The same curve-fitting procedure is applied to estimate the long-term resurrection rate of half-century discards.

10. If correct, my optimistic predictions of the first asymptote at $p_1 < .05$ and the second at $p_2 < .01$ would warrant acceptance of ensconcement/discard as a proxy for Peirce's ultimate consensus.
11. The candidate predictors are combined in a discriminant function (initially linear, as a first approximation), and their statistical weights reflect their relative importance.
12. Deleting the ensconcement/discard criterion, a factor analysis is conducted of the 55 pairwise correlations of the 11 indicators. It is predicted that the first principal component accounts for almost all of the system's variance. The loadings of the indicators measure their validity with respect to that prominent latent factor.
13. Deleting the ensconcement/discard criterion, we conduct a taxometric analysis of the indicators to ascertain whether a latent class (taxon, category) manifests itself in the pattern of indicator correlations. Finding that it does, we assign taxonic validities to the indicators in terms of the separation each achieves between the inferred taxon and the complement class. For each scientific domain the taxon base rate *P* is estimated.
14. Several predictions are made as to relations among these statistical findings:
    (a) The profile of discriminant weights will closely match the profile of loadings on the first principal component.
    (b) The profile of the taxon separations taxometrically inferred will closely match findings from the discriminant function and the factor analysis.
    (c) The taxon base rates in various empirical domains will closely match the directly observed proportions of ensconced theories among theories per domain.
15. An index (VINDEX) of content-matching between various theories at their mid-life and the ensconced theory of a domain will have a high correlation with theories' indicator composites.

The coherence of the above findings, if as strong as I anticipate, warrants a minimalist inference that there is a factor underlying and explaining these relationships, some latent attribute of theories in which they vary widely, involving two strongly separated latent distributions between a taxon and a complement class, such that the taxon corresponds to the manifest ensconcement and the complement class to manifest discard. The asymptotic findings (quite independent of these relationships) assure us that this factor is what underlies ultimate fate. This factor corresponds closely to VINDEX, a numerification of competing theories' content resemblance to the ensconced theory.

A consistent instrumentalist or Peircean pragmaticist should rejoice in these results, the procedures appraising theories satisfactorily given purely instrumentalist aims. The scientific realist does not dispute this, but takes the further step of identifying Peircean consensus with objective truth or high

verisimilitude. The warrant for taking this step is a twofold armchair argument. First, we have theoretical considerations as to why one might rationally expect each of the eleven indicators to be a truth correlate (Meehl [2002]). Second, we have logical arguments showing why false theories should be asymptotically detectable in the long run. If we are quasi-certain that a false theory cannot ultimately survive and that half-century ensconcement is a valid proxy for ultimate survival, then, by Excluded Middle, we can be quasi-certain that an ensconced theory is true or has high verisimilitude. If the cliometric research program turns out as I optimistically predict, both the instrumentalist and the scientific realist should be satisfied.

7 Implementation of Cliometric Metatheory

The proposed cliometric research program is formidable and to many will appear grandiose. While most of the first-level work can be done in bits and pieces as doctoral dissertations, it would be necessary to coordinate these by a central planning agency. The justification for such a huge investment of time and money lies in the consequences of the Faust-Meehl actuarial thesis. If that thesis is correct, the benefits to the advancement of science would be extremely great and long-lasting. Improvement in the appraisal of scientific theories could be as important theoretically and technologically as the mapping of the human genome, the atlas of the stars, or the cataloging of species of bacteria. But I am not optimistic about the conversion of a critical mass of historians and philosophers of science to bring about such an investment in the foreseeable future

Without investing in the full-scale program, there are studies of limited scope, involving far fewer scholars, that would be intrinsically valuable as well as helpful in deciding whether the full program is worth serious consideration. For example, one need not collect data concerning the predictive validity of the whole list of theory attributes but could conduct interesting investigations of selected predictors. One could select a fairly small number of mini-theories in a scientific domain and investigate whether, say, parsimony_2 is a better predictor of ensconcement/discard than is, say, numerical precision. Somewhat larger scale, but still achievable with modest investment and some central coordination, would be investigation of the asymptotic conjecture that fifty-year ensconcement is a good proxy for the inaccessible Peircean ultimate consensus. One hundred doctoral dissertations might be sufficient to provide a trustworthy answer to this question; each would involve studying the fate of an ensconced mini-theory versus that of several discarded competitors, something that can be studied from the scientific literature without reference to the predictor attributes. Suppose it were found that for such a sample across several scientific domains the post-ensconcement reversals were $p < .05$ and $p < .01$, as I optimistically predict. This empirical result, when combined with the *a priori* probability argument

that pseudic theories will be nearly certain to fail Peirce's long-term consensus criterion, comes close to solving the perennial problem about scientific realism. Additional suggestions for implementing cliometric metatheory may be found in Meehl ([1992a]).

Department of Psychology
University of Minnesota
Minneapolis, MN 55455–0344
USA

References

Ayer, A. J. [1940]: *The Foundations of Empirical Knowledge*, New York: Macmillan.

Cronbach, L. J., and Meehl, P. E. [1955]: 'Construct validity in psychological tests', *Psychological Bulletin*, **52**, pp. 281–302. Reprinted in P. E. Meehl, *Psychodiagnosis: Selected Papers*, 1973, Minneapolis, MN: University of Minnesota Press, pp. 3–31.

Dawes, R. M. [1979]: 'The robust beauty of improper linear models in decision making', *American Psychologist*, **34**, pp. 571–82.

Dawes, R. M. and Corrigan, B. [1974]: 'Linear models in decision making', *Psychological Bulletin*, **81**, pp. 95–106.

Efron, B. [1979]: 'Bootstrap methods: Another look at the jackknife', *Annals of Statistics*, **7**, pp. 1–26.

Faust, D. [1984]: *The Limits of Scientific Reasoning*, Minneapolis, MN: University of Minnesota Press.

Faust, D. and Meehl, P. E. [1992]: 'Using scientific methods to resolve enduring questions within the history and philosophy of science: some illustrations', *Behavior Therapy*, **23**, pp. 195–211.

Faust, D. and Meehl, P. E. [2002]: 'Using meta-scientific studies to clarify or resolve questions in the philosophy and history of science', *Philosophy of Science*, **69**, pp. S185–S196.

Feyerabend, P. [1970]: 'Consolations for the specialist', in I. Lakatos and A. Musgrave (*eds*.), *Criticism and the Growth of Knowledge*, Cambridge: Cambridge University Press, pp. 197–230.

Fisher, R. A. [1970]: *Statistical Methods for Research Workers* (14th *edn*), Edinburgh: Oliver and Boyd.

Fisher, R. A. [1971]: *The Design of Experiments* (9th *edn*), New York: Hafner.

Glymour, C. [1980]: *Theory and Evidence*, Princeton, NJ: Princeton University Press.

Gorsuch, R. L. [1983]: *Factor Analysis* (2nd *edn*), Hillsdale, NJ: Erlbaum.

Grove, W. M., and Meehl, P. E. [1993]: 'Simple regression-based procedures for taxometric investigations', *Psychological Reports*, **73**, pp. 707–37.

Harlow, L. L., Mulaik, S. A. and Steiger, J. H. (*eds*.). [1997]: *What If There Were No Significance Tests?*, Mahwah, NJ: Erlbaum.

Hotelling, H. [1936]: 'Relations between two sets of variates', *Biometrika*, **28**, pp. 321–77.

Hull, C. L. [1943]: *Principles of Behavior*, New York: Appleton-Century.
Lachenbruch, P. A. [1975]: *Discriminant Analysis*, New York: Hafner.
Lakatos, I. [1970]: 'Falsification and the methodology of scientific research programmes', in I. Lakatos and A. Musgrave (*eds*.), *Criticism and the Growth of Knowledge*, Cambridge: Cambridge University Press, pp. 91–195. Reprinted in J. Worrall and G. Currie (*eds*.),1978, *Imre Lakatos: Philosophical Papers. Vol. I: The Methodology of Scientific Research Programmes*, New York: Cambridge University Press, pp. 8–101.
Lewis, C. I. [1929]: *Mind and the World-Order*, New York: Scribner's.
MacCorquodale, K., and Meehl, P. E. [1954]: 'E. C. Tolman', in W. K. Estes et al., *Modern Learning Theory*, New York: Appleton-Century-Crofts, pp. 177–266.
Meehl, P. E. [1973]: 'MAXCOV-HITMAX: a taxonomic search method for loose genetic syndromes', in Meehl, *Psychodiagnosis: Selected Papers*, Minneapolis, MN: University of Minnesota Press, pp. 200–24.
Meehl, P. E. [1977]: 'Specific etiology and other forms of strong influence: some quantitative meanings', *Journal of Medicine and Philosophy*, **2**, pp. 33–53.
Meehl, P. E. [1978]: 'Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology', *Journal of Consulting and Clinical Psychology*, **46**, pp. 806–34.
Meehl, P. E. [1983]: 'Subjectivity in psychoanalytic inference: the nagging persistence of Wilhelm Fliess's Achensee question', in J. Earman (*ed*.), *Minnesota Studies in the Philosophy of Science: Vol. X, Testing Scientific Theories*, Minneapolis, MN: University of Minnesota Press, pp. 349–411.
Meehl, P. E. [1990a]: 'Appraising and amending theories: the strategy of Lakatosian defense and two principles that warrant using it', *Psychological Inquiry*, **1**, pp. 108–41, 173–80.
Meehl, P. E. [1990b]: '*Corroboration and Verisimilitude: against Lakatos' "Sheer Leap of Faith"* ' (Working Paper, MCPS–90–01). Minneapolis, MN: University of Minnesota, Center for Philosophy of Science.
Meehl, P. E. [1990c]: 'Why summaries of research on psychological theories are often uninterpretable', *Psychological Reports*, **66**, pp. 195–244. Also in R. E. Snow and D. Wiley (*eds*.), 1991, *Improving Inquiry in Social Science: A Volume in Honor of Lee J. Cronbach*, Hillsdale, NJ: Lawrence Erlbaum, pp. 13–59.
Meehl, P. E. [1992a]: 'Cliometric metatheory: the actuarial approach to empirical, history-based philosophy of science', *Psychological Reports*, **71**, pp. 339–467.
Meehl, P. E. [1992b]: 'The miracle argument for realism: an important lesson to be learned by generalizing from Carrier's counter-examples', *Studies in History and Philosophy of Science*, **23**, pp. 267–82.
Meehl, P. E. [1993]: 'Four queries about factor reality', *History and Philosophy of Psychology Bulletin*, **5** (No. 2), pp. 4–5.
Meehl, P. E. [1995]: 'Bootstraps taxometrics: solving the classification problem in psychopathology', *American Psychologist*, **50**, pp. 266–75.
Meehl, P. E. [1996]: *Clinical versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*, Northvale, NJ: Jason Aronson. (Original publication 1954)
Meehl, P. E. [1999a]: 'Clarifications about taxometric method', *Journal of Applied and Preventive Psychology*, **8**, pp. 165–74.

Meehl, P. E. [1999b]: 'How to weight scientists' probabilities is not a big problem: Comment on Barnes', *British Journal for the Philosophy of Science*, **50**, 283–95.

Meehl, P. E. [2002]: 'Cliometric metatheory II: Criteria scientists use in theory appraisal and why it is rational to do so', *Psychological Reports*, **91**, 339–404.

Meehl, P. E. and Golden, R. [1982]: 'Taxometric methods', in P. Kendall and J. Butcher (*eds*.), *Handbook of Research Methods in Clinical Psychology*, New York: Wiley, pp. 127–81.

Meehl, P. E. and Waller, N. G. [2002]: 'The path analysis controversy: A new statistical approach to strong appraisal of verisimilitude', *Psychological Methods*, **7**, pp. 283–300.

Meehl, P. E. and Yonce, L. J. [1994]: 'Taxometric analysis: I. Detecting taxonicity with two quantitative indicators using means above and below a sliding cut (MAMBAC procedure)', *Psychological Reports*, **74**, 1059–274.

Meehl, P. E. and Yonce, L. J. [1996]: 'Taxometric analysis: II. Detecting taxonicity using covariance of two quantitative indicators in successive intervals of a third indicator (MAXCOV procedure)', *Psychological Reports*, **78**, 1091–227.

~~Miller, D. W. [1975]: 'The measure of all things', in G. Maxwell and R. L. Anderson, Jr. (*eds*.), *Minnesota Studies in the Philosophy of Science: Vol. VI, Induction, Probability, and Confirmation*, Minneapolis, MN: University of Minnesota Press, pp. 350–66.~~ [incorrect reference in original]

Miller, D. W. [1975]: 'The accuracy of predictions', *Synthese*, **30**, pp. 159–91.

Morrison, D. E. and Henkel, R. E. (Eds.). [1970]: *The significance test controversy*, Chicago: Aldine.

Morrison, D. F. [1990]: *Multivariate Statistical Methods* (3rd *edn*), New York: McGraw-Hill.

Niiniluoto, I. [1998]: 'Verisimilitude: The third period', *British Journal for the Philosophy of Science*, **49**, 1–29.

Pap, A. [1958]: *Semantics and Necessary Truth*, New Haven, CT: Yale University Press.

Peirce, C. S. [1878/1986]: 'How to make our ideas clear', in C. J. W. Kloesel (*ed*.), *Writings of Charles S. Peirce* Vol. 3, Bloomington, IN: Indiana University Press, pp. 257–76. (Originally published in *Popular Science Monthly*, **12**, pp. 286–302)

Waller, N. G. and Meehl, P. E. [1998]: *Multivariate Taxometric Procedures: Distinguishing Types from Continua*, Newbury Park, CA: Sage.

Waller, N. G. and Meehl, P. E. [2002]: 'Risky tests, verisimilitude, and path analysis', *Psychological Methods*, **7**, 323–37.

pdf by LJY October 2010