Psychological Reports, 2002, *91*, 339-404.

#180

Cliometric Metatheory II: Criteria Scientists Use in Theory Appraisal and Why It Is Rational to Do So[1, 2]

Paul E. Meehl

University of Minnesota

Summary—Definitive tests of theories are often impossible in the life sciences because auxiliary assumptions are problematic. In the appraisal of competing theories, history of science shows that scientists use various theory characteristics such as aspects of parsimony, the number, qualitative diversity, novelty, and numerical precision of facts derived, number of misderived facts, and reducibility relations to other accepted theories. Statistical arguments are offered to show why, given minimal assumptions about the world and the mind, many of these attributes are expectable correlates of verisimilitude. A statistical composite of these attributes could provide an actuarial basis for theory appraisal (cliometric metatheory).

CONTENTS

Many contemporary philosophers of science (e.g., Quine, 1969; Sneed, 1976) prefer to view their subject (except for formal logic) as an empirical discipline, a branch of behavioral science, say, social or cognitive psychology. From this viewpoint, while one may aim to provide a rational reconstruction of scientific

[1] I am grateful to Leslie J. Yonce for assistance with this article, to Dean Keith Simonton for helpful comments on an earlier version of this material, and to David Faust and William M. Grove for clarifying conversations.

[2] [Related publications on cliometric metatheory include Meehl, 1990a, 1992a, 1992c, and 2004.—LJY]

theorizing, the starting point has an empirical basis: (a) episodes in the history of science, (b) current practice of scientists, (c) experimental psychology of perception, memory, problem-solving, and the like. Philosophy of science is thus taken to be *metatheory*, the empirical theory of scientific theorizing. Cliometric metatheory (as explained in Meehl, 1992a) considers philosophy of science as the empirical theory of scientific theorizing. Its database consists of episodes in the history of science, sampled actuarially and analyzed by sophisticated psychometric methods.

Exploration of the historical basis usually proceeds by case studies, to which I offer no objection; but that method needs an important qualification. A case study—say, of the rise and decline of a theory, the misinterpretation of a crucial experiment, or the scientific community's resistance to discovery (Barber, 1961)—can shed light on how science works and can provide fruitful suggestions for appraising and amending metatheory. Second, when carefully done as to the facts and their first-level interpretation, a case study can refute a metatheoretical conjecture. But when the case study method is employed not to suggest or to refute but to *confirm* a metatheoretical generalization, we immediately confront the well known problem of generalizability. We know that anecdotal impressions (honorifically labeled "clinical experience" when formed by PhDs and MDs) are frequently misleading. The whole history of human superstition and of horrors such as witchcraft persecutions attests to the untrustworthiness of anecdotal impressions as sufficient basis for either theoretical understanding or pragmatic efficacy. Physicians routinely and universally employed such useless and harmful procedures as bleeding, purging, and blistering, justified on the basis of clinical experience. The 1899 *Merck Manual* (1999) lists over 800 drugs, most of which were inefficacious.

Writing for readers of this journal, I hardly need to prove that anecdotal impressions are not a trustworthy source of knowledge (Meehl, 1997a).[3] Considering the sources of error in clinical judgment (see Meehl, 1992a, Table 1, pp. 353-354 for a list of several dozen), we find that identical or closely similar factors must operate to impair both a scientist's impressionistic appraisal of scientific theories and the philosopher (or historian) of science's evaluation of metatheory. The proper way to employ history of science in reaching valid generalizations about how scientists proceed—and the more daunting task of forming metatheoretical prescriptions of how we *should* proceed—is by appropriate kinds of

[3] Misreading my *Clinical versus Statistical Prediction* (Meehl, 1954/1996) led some clinicians to perceive me as a generic denigrator of clinical experience, numerous passages in that book to the contrary notwithstanding. I earned part of my living for half a century in the practice of psychotherapy (psychoanalytic and rational emotive), relying almost wholly on my clinical experience and that of Freud and Ellis, neither of whom did experiments or computed statistics, with the notable exception of Ellis' path-breaking comparison of himself in three modes (1957).

random and representative sampling of the fact domain, that is, by a statistical approach to scientific literature. The Faust-Meehl strong actuarial thesis states this position: "*Metatheoretical research should (1) make actuarial summaries of the properties and fates of scientific theories based on random sampling of episodes from the history of science and (2) apply formal analytic methods (e.g., psychometrics) to appraise metatheoretical conjectures*" (Meehl, 1992c; see Faust, 1984; Faust & Meehl, 1992, 2002).

Some younger philosophers of science, especially the so-called "positivist-bashers," seem to think that an empirical conception of metatheory implies that the traditional questions of the philosopher ought not (cannot?) be asked; or that, if asked properly, the conceptual tools for answering them cannot include the conventional ones. This is a mistake. The traditional questions about definition (explication, clarification) of theoretical concepts and about justification (intrinsic merits and evidentiary support) of theories are just as meaningful today as in the days of Descartes, Locke, and Hume. Nor can we dispense with the four traditional tools for such explication and justification: logic, semiotic (syntactics, semantics, pragmatics), probability theory, and what may be called "armchair epistemology" (Meehl, 1992a). Scientists, unlike stones and daffodils, engage in cognitive activities. They assert sentences reporting observations. They subsume, generalize, deduce, defend, clarify, explain, define, deny, refute, and speculate. They derive mathematical theorems and compute statistics. It is not possible even to *describe* scientific behavior without the use of distinctively philosophical concepts (proof, consistency, contradiction, derivation, generalization, and the like); and it is *a fortiori* not possible to *explain* the results of scientific reasoning, e.g., the long-term success or failure of a theory, without using the four traditional tools.[4]

Imagine an economist who is interested in understanding business failures. He collects a batch of bankruptcy cases and writes up case studies of them. He finds, for example, that the president of Widget Co. had an IQ of 90 which, while normal, is below what is required to be a competent business executive in our society. The CEO of Dwidget Co. was an alcoholic who made afternoon business decisions after a four-martini lunch. The plant manager of International Silk-lined Casket Corporation was embroiled in a prolonged acrimonious divorce proceeding, working on an ulcer, and had severe insomnia. These are the sorts of examples our economist submits to a scientific journal in an article purporting to explain business failures. No journal editor, economist, or psychologist would

[4] I do not refer to the obvious necessity that the metatheorist reason logically, accept facts, define terms, calculate, and the like. That goes without saying. My point is that the metatheorist cannot avoid employing meta-talk about the scientist's rational (and irrational!) cognitive processes in the metatheoretical *causal* account. Hence, metatheory is substantively different from first-level theory in geology or biology because stones and daffodils do not think but scientists do.

accept such a paper. Why not? Not because stupidity, drunkenness, and anxious preoccupation cannot serve as "explainers," as crucial initial links in the causal chain leading to bankruptcy. The trouble is absence of the intermediate links in the causal chain. The psychology of these three executives is the first link in the chain, but the intervening events are *unsound decisions*, episodes of poor business judgment consequent upon the three personality deficiencies; and the notion of an unsound business decision involves intrinsically economic concepts. If we deprive ourselves of such distinctively economic concepts as marginal cost, intersection of supply and demand curves, economies of scale, consumer brand loyalty, the government's fiscal and monetary policy, the reserve capital of a competitor, and the like, we cannot even classify these poor decisions as wise or unwise.

This article presents criteria to be used in cliometric theory appraisal and shows why each may be expected to correlate with the long-term success or verisimilitude of a theory.

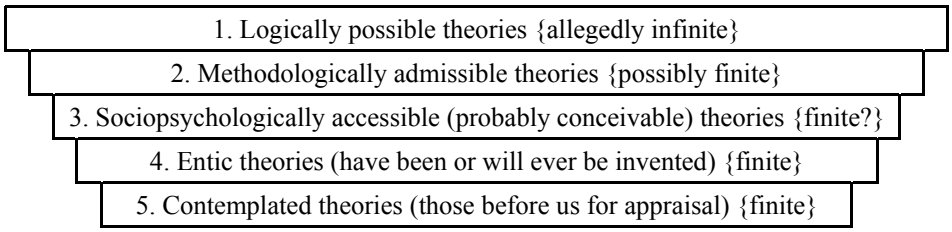
CLASSES OF THEORIES FOR CLIOMETRIC APPRAISAL

There is a problem presented by the logician's contention (they take it as "obvious," but I have yet to see the formal proof of a theorem) that for any finite collection of empirical facts there is an infinite number of logically possible explanatory theories. It is not clear how this alleged metatheoretical truism bears on empirical metatheory, if at all; no scientist of my acquaintance has ever worried about this infinity of abstract possibilities. In reality, of course, the count of plausible theories for most fact domains is rather limited. One indicator that a science is in a primitive state is a proliferation of theories.[5] To carry out the cliometric research program, we treat all of our classes, whether of facts or theories, as an insurance actuary does, that is, as being finite, however large. In order to ensure this, think of nested classes (summarized in Fig. 1) of theories as follows:

1. *Logically possible theories*.—These allegedly are infinite in number.

2. *Methodologically admissible theories*.—This excludes almost all theories in (1). I cannot prove that this class is finite, although I believe it is, as do most scientists. Scientists impose methodological constraints on theories that go beyond purely logical requirements (consistency, nonredundancy, definitions). Some of these constraints are formal, others involve the *kind* of theoretical entity or process postulated, others involve compatibility with background knowledge. An important variant of the latter stems from Comte's Pyramid of Sciences, which strongly constrains theorizing at one level by reference to ensconced theories at levels above and below. For example, a theory of mitosis in which the spindle fibers were postulated to be fine platinum wires would be summarily

[5] My colleague, Will Grove, and I, during a 10-min. break in our seminar, came up from memory with 45 theories of schizophrenia, and I have read somewhere that there are over 100 such.



A given theory is specified by postulates: $P_1, P_2, \ldots P_n$.
T_T : the "perfect," complete and true theory, asserts *all* of the postulates correctly.
The theories we deal with may be
Apseudic (assert some true postulates and not assert any false postulates)
or Pseudic (assert one or more false postulates).

		Postulate Coverage	
		Complete	Incomplete
Postulates Asserted	True	T_T Apseudic	Apseudic
	False	Pseudic	Pseudic

Fig. 1. Nested classes of theory populations and kinds of theories.

rejected by every biologist. The conventionally admitted thesis of "underdetermination of theories by facts," however numerous, qualitatively diverse, and numerically precise, seems pretty clearly incorrect for developed sciences, and not all philosophers subscribe to it (see Boyd, 1973; Glymour, 1980; Wilson, 1980). *Example*: Given the available facts, no sane person would attempt to concoct a theory of the gene that did not involve the double helix of ribose and phosphate radical, lacking ordered triads of adenine, guanine, cytosine, and thymine corresponding to the 20 amino acids. That theory is as firmly ensconced as the existence of Singapore. I view this disparity between a (purported) logical truism and scientific practice to be one of the most important and badly neglected metatheoretical puzzles. For a few examples of methodological constraints see Meehl (1990b). I am confident that content analysis of scientific texts criticizing theories as to threshold admissibility would yield numerous other (largely tacit) exclusionary principles.

3. *Sociopsychologically accessible theories*.—This is the subset of admissible theories for which there is a nonnegligible probability of some scientist conceiving them, even if nobody actually does. I would adopt a convention that we count a theory as accessible if it has a probability $p > 10^{-4}$, that is, the number that

Buffon thought negligible, rationally treatable as if it were zero (Todhunter, 1865; Keynes, 1921; Gillies, 1973; Meehl, 1992a). Arguably, an equally "reasonable" convention for accessibility would be closer to that of the applied statistician, that a theory with invention probability is $p > .01$ or $p > .05$ is accessible. If the latter convention renders the class finite and Buffon's allows it to be infinite, we use the stipulation that gives us a finite set.

4. *Entic theories*.—An entic theory is one which has been or will be invented before the sun burns out or the human race destroys itself. These are theories that are (logician's tenseless 'are') in fact concocted. This class is obviously finite.

5. *Theories in contemplation*.—This is the important category for the working scientist, being the theories that are before us for appraisal, and it is finite.

Because we can be sure that classes (4) and (5) are finite and conjecture that (2) and (3) are, we do not worry about technical problems that arise from probabilities involving countable or uncountable infinities. Some have suggested that the total number of theories in science is too small to treat statistically. I do not understand that argument, since it is a finite population, and sampling error in the usual sense can therefore be theoretically avoided by considering all of them. However, the notion that the number of theories in empirical science is small is clearly erroneous. I suspect it comes from focusing excessively on the grand theories that we like to use as favorite and exciting examples, such as relativity, Newton's mechanics, optics, thermodynamics, quantum mechanics, the theory of evolution, the germ theory of disease, operant behavior theory, Keynesian economics, and so on. This gives a very misleading impression as to the number of theories. It is easy to see that even in a fairly restrictive domain such as, say, human physiology, theories number in the thousands. They are *minitheories* rather than grand theories, but they are theories in the fullest sense and deserve detailed logical, epistemological, and statistical analysis. They are often very complex, requiring many years and thousands of research papers to develop and confirm. For example, it took 75 years—three generations of scientists—to complete the theory of pernicious anemia, from the 1920s when the liver treatment was introduced to the 1990s when the codon sequence underlying synthesis of the missing intrinsic factor was finally determined. Genetics of human disease may seem a narrow domain, hence yielding only a few theories for taxometric study, but McKusick's atlas [1998, *Mendelian inheritance in man: a catalog of human genes and genetic disorders*. (12th ed.) Baltimore, MD; Johns Hopkins Univer. Press. Available: http://www.ncbi.nlm.nih.gov/Omim/] lists over 9000 syndromes, each involving complex compositional and causal relations among theoretical entities, processes, and states.

An additional classification of theories that we must keep in mind concerns the extent to which they correspond to the true postulates of a domain. For any given domain there will be one "perfect" theory, T_T, the theory that is both true

and complete; it is the theory that *derives all true* operational statements. Theories that we work with, however, are imperfect in one or more ways. They are generally incomplete; they assert some but not all of the true postulates for the domain. Because they are incomplete, they may also derive false predictions, and they will always leave some operational statements undecided. No psychological theory derives the outcome, correct or incorrect, of every conceivable experiment. Generally, I will call a theory *apseudic* (from 'pseudo' plus the alpha privative) if it is incomplete (or possibly complete) but does not assert any false postulates.[6] If a theory asserts at least one false postulate, I will call it *pseudic*. In the social and life sciences, apseudic theories are probably the best we can hope for.

CRITERIA FOR THEORY APPRAISAL

Although many have offered lists of desirable theoretical properties and criteria for evaluating theories (see, e.g., Feigl, 1929, pp. 131–137; Cohen & Nagel, 1934, pp. 207–215; Margenau, 1950, pp. 81–121; Frank, 1954; Copi, 1961, pp. 426–433; Popper, 1962, pp. 231–233; Hempel, 1966, pp. 33–46; Schaffner, 1970, pp. 318–330; Kordig, 1971a, 1971b, 1978; Kuhn, 1977, pp. 320ff; Shapere, 1977; Newton-Smith, 1981, pp. 226–232; Laudan, 1984; Watkins, 1984, p. 130ff *et passim;* Dauer, 1989; Faust & Meehl, 1992; 1950, pp. 38–41 [1981, pp. 196–200]; Meehl, 1992a, pp. 379–380, 406; Thagard, 1992), the literature of metatheory and history of science is surprisingly thin as to criteria of theory appraisal. Most articles and books mention only two or three—the only one they all (except Popperians) mention is parsimony—and few lists are longer than four or five. This is puzzling, because there is hardly any contribution that philosophers and historians of science could make to the working scientist that would be more valuable than how better to appraise theories. I have had to rely more on my introspections and my observation of other scientists than on the metatheoretical literature.[7]

---

[6] One might think that missing postulates could render an apseudic theory pseudic. For instance, if a seemingly apseudic theory asserts that an observational variable is a function of theoretical variables θ_1, θ_2, and θ_3, leaving out a variable θ_4 that would be known in T_T, then it has asserted a bridge law falsely. But the bridge law is itself a postulate (the "operational" linkage), hence the theory is pseudic. See Meehl (in preparation [2004]) for further discussion of such refinements.

[7] Although the empirical source of my list admittedly is anecdotal rather than cliometric, I think I am in a somewhat better position than most to arrive at such generalizations, having worked in several different areas of psychology (animal learning, psychometrics, behavior genetics, psychoanalysis, clinical prediction, interview assessment, personality theory, forensic psychology) and having engaged in interdisciplinary research or advanced seminar instruction with geneticists, neurologists, psychiatrists, sociologists, political scientists, and statisticians. For the physical sciences I rely on extensive reading and 20 years of attendance at colloquia held by the University of Minnesota History of Science and Technology program.

TABLE 1

ATTRIBUTES USED BY SCIENTISTS IN THEORY APPRAISAL.

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
Additional criteria:
Initial plausibility
Rigor of theoretical derivations
Confidence in the auxiliaries in observational testing
Deductive fertility, fruitfulness
Technological power
Computational ease
Beauty, depth, elegance

Table 1 lists criteria that scientists sometimes employ in appraising theories. There is no agreement as to the relative importance that should be rationally assigned to each of them, and here is where the Faust-Meehl actuarial thesis comes into play, challenging the conventional nonstatistical case studies approach.

Contemplating my candidate list of indicators, one might ask—say, of 'parsimony' (there are four kinds!)—why pay attention to it? Does anybody claim that parsimony is a litmus test for ultimate survival of a theory (Charles Sanders Peirce's pragmaticist definition of truth[8]) or objective verisimilitude (a realist's definition of truth)? Of course not. When comparing two theories parsimony may favor one, and another attribute, say, numerical precision, might favor the other. Does parsimony trump numerical precision? Or does parsimony trump all of the other criteria? Nobody would maintain that, either. Thus, since parsimony is not

[8] In treating of long-term survival, I am not adopting Peirce's famous definition of truth as "the opinion which is fated to be ultimately agreed to by all who investigate… and the object represented in this opinion is the real" (Peirce, 1878/1986). I am a scientific realist rather than a pragmaticist. I here treat survival as a proxy for truth and have elsewhere presented reasons for why it is rational to take it as a valid proxy (Meehl, in preparation [2002]). The pragmatist avoids this last step and can dispense with the ensuing arguments, since long-term survival—the theory having solved its problems (Laudan, 1984)—is for him not a proxy but the criterion, the defined aim.

a guarantee of truth nor its absence a guarantee of failure, nor is it an attribute that trumps all of the others, the obvious question is, "Why should we pay any attention to it?" There is only one possible answer to that query: If we pay attention to a theory attribute, it is because we hold (assume? hope?) that it is a *correlate* of survival or verisimilitude.

What sort of claim is that? It is an *empirical* claim. We may have rational grounds for supposing that parsimony is a correlate of ultimate survival; but however sound or unsound those may be, the semantic *content* of the claim is about a matter of (future) fact. If it should turn out after centuries of metatheoretical research that, contrary to our armchair expectations, parsimony is not correlated with long-term theory survival, we would have to abandon our generalization and our habitual reliance on it in theory appraisal. We would also have a problem of philosophical analysis on our hands, namely, to show why the plausibility arguments for parsimony are unsound.

THEORY PROPERTIES AND RELATIONS AS EXPECTABLE TRUTH CORRELATES

There are plausibility arguments for expecting the first 11 criteria in Table 1 to be statistically correlated with theory truth. The additional indicators are not examined here, for one or more of the following reasons: They are more rarely mentioned, and some reject them; I doubt their validity and am unable to construct strong theoretical arguments for their being truth-correlated; some may merely be composites or consequences of the first eleven; I do not readily see how to quantify them. If they do have merit and contribute incremental validity, this should be ascertained at a later stage of the cliometric program.

While the statistical arguments I give for these criteria are, I trust, formally valid, they are nevertheless only plausibility arguments because they depend upon general assumptions about the world and the human mind that are not indubitable. With respect to Laudan's scientific principles (1984), the question has been raised (Schmaus, 1996) whether they should be considered analytic or empirical. Whatever may be true of Laudan's, there can be no doubt that my list is empirical. All derivations via the formalism of the probability calculus must begin with (contingent) probability numbers, since that calculus can never be more than a way of inferring some probabilities from others.

Consider the "worst case" of scientific knowledge as pressed by super skeptics, namely, that the human mind is incapable of formulating true theories about anything. If that were so—no one has succeeded in offering a rigorous guarantee that it could not be so, and some clever, scientifically informed persons profess to believe it—then any logical or mathematical derivation purporting to show that some property or relation of theories is a truth correlate must be unsound, it being a mathematical truism about correlation that x and y cannot be correlated unless they both vary. Hence nothing can be a truth correlate if all theories are false. A claim that the class of theories invented by human scientists

that have some property Q has a higher truth frequency than the class of theories that lack Q is a *factual* claim, however one purports to reach it and whether or not such a thing can be reliably known. If I were to derive such a factual claim from something other than another factual claim, I would either be getting it from a tautology, which we know is impossible, or from a synthetic *a priori* truth, which doesn't exist. In all of the derivations offered here, the empirical assumption (better, conjecture) will usually be so obvious as hardly needing statement.[9]

Despite this disclaimer, one can prove that if measuring instruments are sufficiently precise, the falsity of the conjunction of a theory and its testing auxiliaries will be detected with probability → 1 as the fact domain is sampled indefinitely (Meehl, in preparation [2004]).

In considering my candidate list, certain *meta*-meta-principles should be kept in mind, which I list briefly without arguing them:

(1) The *properties* of theories are formal, e.g., mathematical "simplicity," and material, e.g., what kinds of entities are postulated; the *relations* of theories include their relations to facts, and their relations to other theories, e.g., partial theoretical reducibility of neuron conduction to microanatomy of the cell and the physical chemistry of semi-permeable membranes. I shall employ 'characteristic' to avoid clumsy repetition of 'property or relation.'

(2) Characteristics are examined one at a time, separately, and derivations of their truth correlation are considered apart from whether some other characteristic, perhaps negatively correlated with the one being considered, can "get in the way." I expect this piecewise mode of consideration to trouble some philosophers, especially logicians, but the cliometric approach bypasses it deliberately with a clear conscience. The statistics of multiple regression and discriminant function are mathematically built to take such relations into account. The non-statistical reader might be happier if, for each theory characteristic, I said, "*ceteris paribus*, property Q is a truth correlate"; but since *cetera* are here never (literally!) *paria*, that counterfactual meta-remark would be more soothing than clarifying. *In cliometric metatheory one's mental set must be consistently statistical*.

(3) Should a plausibility argument be unsound, either because the primary empirical supposition is false or, as in some of my derivations, a plausible

[9] If the reader is troubled because much of this argument has the appearance of old-fashioned 'first philosophy,' my answer is that old-fashioned first philosophy, whether practiced by British empiricists, or Continental rationalists, or logical positivists, always did begin with certain factual premises about the human mind in relation to the world. "Cognitive science," which collects statistics about beliefs and does experiments on problem-solving, was not the beginning of a basic empiricism in epistemology.

auxiliary conjecture plays a crucial role, no long-term harm is done. All that is claimed for any of these derivations is that they are plausible enough to justify including the characteristic in the candidate list of predictors of long-term theory ensconcement, which is in turn taken as our best available indicator of truth.

(4) While my frame of reference is that of scientific realism, the consistent instrumentalist should have little trouble making the necesssary translations. In a sense, the cliometric approach is more easily instrumentalist than realist, because the idea of Peircean ultimate consensus is closer to instrumentalist thinking about truth than is my taking ensconcement as a truth proxy.

(5) Much of the treatment is idealized, for which I make no apologies. Some philosophers seem to believe that meta-theoretical discourse cannot involve conceptual or numerical approximations, although first-level scientific theoretical discourse routinely does so. Why should we assume that all metatheoretical talk can be precise, and when it is not precise it must be rejected? *Prima facie*, one might expect idealizations, approximations, and incompletenesses to be even greater and harder to eliminate in a metatheory that calls itself naturalized than in the other kind, or than in first-level (object language) theories.

(6) Considering the finite sets of accessible, entic, or contemplated theories enables us to derive the desired "forward-looking" truth correlation, rather than relying on the "otherwise remarkable coincidence," backward-looking indirect argument. I do not disagree with that latter argument or think it unhealthy for scientists or philosophers to use it, but since it has been attacked as not showing *quite* what one wants to show, I aim throughout to achieve the forward-looking correlation. We wish to say, for a class of theories dichotomized into a subclass having attribute Q and a subclass lacking Q that the truth frequency of the former exceeds that of the latter; theories with Q are *more probable*.

Given that aim, we must keep in mind a simple algebraic fact about the fourfold table in statistics, without which the reasoning would be suspect. Considering two properties A and B, which the members of a class of entities, abstract or physical, may possess or lack, there is a simple relationship among conditional probabilities,

$$\text{If } P(A|B) > P(A) > P(A|\bar{B})$$
$$\text{then } P(B|A) > P(B) > P(B|\bar{A}).$$

It is crucial to understand that this necessarily holds for the numbers in any fourfold table of frequencies. The inference from "True theories have a higher incidence of property Q than do false theories" to "Theories having the property Q have a higher truth-frequency than theories that lack Q" is direct. It does not

depend on a covert Bayesian assumption about priors, nor can it be refuted by any adverse Bayesian considerations. Of course, if we tried to infer directly that a theory having property Q is *more probable than not*, we would be in trouble without additional assumptions. If we wanted to establish a "correct" metric of the probabilities, we would be in trouble. If we alleged that some other property Q_2 could not countervail the presence of Q_1 as an indicator of truth-frequency, we would be in trouble. But we are not going to do any of those three things. Whether they are do-able, I do not consider. In the plausibility arguments that follow I shall refer to the above relations as *correlational symmetry*.

Without having done an actuarial tally, I say confidently that in the various lists of desirable theory properties set out by scientists, philosophers, and historians of science, the one which is almost never absent, even in very short lists, is parsimony.[10] There are at least four distinguishable sorts of parsimony. Whether two or more of these can be viewed as equivalent in some deep epistemological sense I do not discuss, and I shall examine their alleged relationship to truth frequency separately.

Parsimony$_1$: *Simplest curve*.—Courses in statistics, and even those labeled more generally "quantitative methods," taken by students of the life sciences are often inadequate with respect to the curve fitting problem. One is told that "of course, the simplest curve is to be preferred," without clarification as to whether this preference is an instrumental one (ease of finding and applying) or one theoretically motivated in the sense of the simplest curve being more likely to be the true curve or closer to the true curve. Lacking both a rigorous definition of 'simplicity' in this context and a mathematical proof that, however defined, it is a truth correlate, the student (and subsequent practicing scientist) hardly has a clue about how to proceed.

In college algebra, one learns that any single-valued discrete function, that is, a set of k points ($x \rightarrow y$), can always be exactly fitted by a polynomial of $(k - 1)$ degree. It is rightfully—but not helpfully—pointed out, "Of course, no empirical scientist would deal with a large set of points in that way." The folly of such a proceeding is intuitively clear from the consideration that between various points the resulting curve may wander erratically in high mountains and deep valleys, so that if one collects additional data points they will not be well-fitted. Improving the advice by saying, "One should prefer simpler curves to more complex ones, provided that the simpler ones do an adequate job fitting the empirical data," is a

[10] Among some psychologists, this methodological preference is even called a "law" and is then, by some behaviorists, conflated with a related but distinct prescription about theorizing concerning infrahuman animals known as Lloyd Morgan's Canon, often used as a polemical club with which to beat one's theoretical opponents. One maverick animal psychologist called it "the postulate of impoverished reality."

step in the right direction but still does not tell us just what specifically to do. The most common advice I have seen in the usual brief treatments of curve fitting for social and biological scientists is to select a curve type, a function form, e.g., straight line, growth function, parabola, on the basis of theory; given that initial choice, it is then a fairly straightforward problem, e.g., least squares, to assign parameters so as to achieve the best fit of which that function form is capable. This advice, which in advanced states of some sciences may be appropriate, is not helpful if we are using the parsimony principle together with the goodness of fit of a curve for the purpose of appraising the theory. And, whichever direction we are arguing, some satisfactory general definition of 'parsimony' as regards curves in relation to data points is imperative.

When the adequacy of a fitted curve is seen to involve a quantitative compromise between overfitting and underfitting, so that the choice of function form is not made initially but made in the light of the results of optimizing parameters for whatever function forms are admissible (for whatever reason), one desires to characterize a "good curve" and ideally a "best curve" with respect to a given data set as some optimizing function of goodness of fit and mathematical simplicity. Since we maximize fit by optimizing choice of parameters for the set of curves constituting a family (curve types, function forms), one looks for a choice function that takes account of the sum of squares of deviations of the fitted curves and somehow counterbalances the goodness of fit with a function of the number of freely adjustable parameters. Forster and Sober (1994), relying on a fundamental theorem due to Akaike (1973), develop this argument in detail and present the derived formula, which expresses the closeness of the fitted curve to the inferred true curve and then, relying on this closeness index, probabilifies the fitted curve with respect to truth. Their epistemological application of the index has been criticized by Kukla (1995) and DeVito (1997), to the first of which Forster (1995) has replied. Pending resolution of this controversy, I shall assume that this, or some amended index, of closeness to the true curve, does the job.

Given an appropriate index of curve fitting parsimony, a distinction must be made. Employing the word 'theory' broadly (as most philosophers use it), we include as a limiting case of theory law-like statements in the observation language involving functional relations between properties and events (without the postulation of unobserved theoretical entities). For example, on a positivistic interpretation of *fields* in which one says, "All I mean by the electrostatic field strength at this point is what acceleration a charged particle will undergo there," such a minimalist attribution is considered correct if the dispositional statement is true and the function from which this disposition is calculated is the empirically correct function. Most of the theoretical postulates of Skinner's (1938) operant

behavior theory are of this sort; similarly are minimally interpreted statements of classical psychometrics (e.g., the first factor exhibited in the matrix of correlations of subtests of an omnibus intelligence test), demand curves in economics, an epidemiological statement of infectious disease spread, and the like. Given a minimalist interpretation that says nothing, or almost nothing, beyond a quantitative formulation of the observation-language dispositions, finding the correct curve is practically indistinguishable from finding the "correct theory."

A more interesting case arises when we have a substantive theory that involves inferred unobserved entities (Reichenbach's *illata* rather than *abstracta*, 1938). Here we have to ask why one should expect curve fitting parsimony in the Forster-Sober sense to be correlated with the truth or verisimilitude of the substantive (structural, compositional, or causal) theory that is not equivalent to the dispositions formulated by the curve but rather entails them, giving rise to the usual problem about the inductive logic of inferred entities. Unfortunately, there are two plausible lines of reasoning that have, at least for me, about equal intuitive appeal, neither of which is rigorously derivable and that will sometimes deliver different answers. A third approach, however, appears to settle the matter.

One line of reasoning starts with the intuitively obvious notion that, considering the class of fitted curves that are more complex, there are more different curves of the same order of complexity, i.e., that have the same value of k = number of adjustable parameters, than there are curves of lower complexity. (The limiting case of this relation is the simplest curve, a straight line with only two adjustable parameters.) If there are function forms in our (how determined?) "admissible" list of curve types, with $k_1 > k_2$, and if there is a one-to-one relation between function forms and the substantive theories that entail them, then the class of substantive theories corresponding to the less parsimonious sets of curves is a larger class. If, as will sometimes be the case, there are alternative competing theories which, despite not being isomorphic, nevertheless are capable of entailing the same function form for a specified dataset, that makes that theory class even larger. For the class of theories that includes the true theory in each fact domain under consideration, the numerator of truth frequency being the number of domains, and the denominator being the number of competing theories over all the domains, then the proportion of true theories (epistemically, the probability of randomly picking the true theory in the set) is smaller for more complex entailed curves than for simpler ones.

Pressing in the opposite direction, a difficulty arises, apart from Kukla's criticism, when we consider curves either that do not differ significantly in the statistician's sense in their adequacy to the data or which, while differing in a statistically significant way, do not differ a great deal. Then the question of

verisimilitude (rather than exact truth) enters the picture, and one can easily think of situations in which a nonmathematical epistemic consideration would countervail advice flowing from the Forster-Sober index. This would happen if a slightly more complex curve, providing a fairly adequate fit, was derivable from a substantive theory but its competitor, perhaps doing slightly better by the Forster-Sober index, was a theoretically unmotivated pure curve-fitting function. Confining ourselves to admissible functions that are analytic, we know that, relying on Taylor's theorem, we can represent any such function by a power series in the input variable. If we approximate by dropping all of the terms in Taylor's expansion involving the third or higher derivatives, we have a parabola and three adjustable parameters. But some other function might do equally well (or even a little worse) from the standpoint of the sum of squares and involve three parameters but a different function form. The departures of this other function from the data points do not occur in the same places, and this other one is theoretically motivated. *Example*: In the Van der Waals correction of the gas law, parameters b and a represent, respectively, the unknown volume occupied by the molecules and the influence of the Van der Waals force in deflecting them when they are close enough together so that the momentum vector normal to the piston head suffers a decrement due to the molecules swerving before they strike it. Van der Waals, writing before the Rutherford atom and treating molecules as his smallest theoretical unit, had no basis for computing these parameters. But the substantive theory entails a certain function form, and it turns out that at very high densities the behavior of the volume-pressure curve deviates from the earlier approximation $PV = RT$ in a somewhat complicated way. Even if the number of adjustable parameters here were the same as in a Taylor-based polynomial fit, some (most?) scientists would tend to view the Van der Waals function as more "complex." So would I, but I do not know how to defend my intuition on that score. With equal numbers of adjustable parameters, I think most of us would prefer the Van der Waals function, even if the fit were not as good. As we know, that correction only takes care of the two idealizations about negligible volume and no attractive forces and still is an approximation because it leaves out a component of the total energy associated with rotational moments. I do not have a way to justify this preference for a theory of "motivated fit" that is perhaps not significantly different from a blind atheoretical polynomial fit. But this and the previous line of thought entitles us to treat curve fitting parsimony as a candidate predictor.

A simple plausibility argument that an *illata* theory that entails a simpler curve is more probably apseudic than one entailing a less simple curve perhaps takes care of the puzzle. Theory T_1 entails observational function $F_1(x, y)$. Theory T_2 entails function $F_2(x, y)$, and F_1 is simpler than F_2 by the Forster-Sober

criterion. Then F_1 is more likely the true function than is F_2. Over domains where such competitive theory-pairs exist, the F_1's have a higher truth-frequency than do the F_2s. For the (larger) subset where F_1 is true, F_2 is incorrect, they being mathematically nonequivalent. Hence for that (larger) subset of pairs, T_2 is pseudic, *modus tollens*. Although F_1 does not entail T_1, if *some* of the T_1's are apseudic, their apseudic rate [> 0] exceeds that of the T_2s [= 0].

Parsimony$_2$: *Economy of postulates*.—"Parsimony" in the sense of making fewer assumptions is the only metatheoretical predicate included in all the lists I have seen compiled by historians, philosophers, or philosophizing scientists (except Popper), and this unusual consensus would by itself warrant its inclusion in our cliometric candidate list. However, a rational basis for this preference for parsimonious theories is rarely stated. The motivation varies over notions of theoretical elegance, pragmatic considerations of ease in manipulation, and application of the theory to technological problems. It does not always involve the belief—almost never argued even when present—that for a given fact domain in which two theories "explain the facts," the more parsimonious theory is more likely to be correct. Since this latter is our focus, plausibility arguments for such a relation are in order.[11]

An intuitive argument is simply, "the more you assert, the riskier." This dictum applies directly when comparing a theory (a . b . c . d) with a "shorter" theory consisting of a proper subset (a . b . c) of the first theory's postulates. But that is not, of course, the usual situation. Comparisons of interest involve two theories, either with totally different postulates, such as (a . b . c . d) versus (e . f . g), or comparing (a . b . c . d) with (a . e . f). Here the usual straightforward consideration of "logical probability" does not do the job. Usually we are not asking whether adding a further postulate is risky; we are comparing different theories, which may or may not have a partial overlap in the postulates asserted.[12]

To see whether it is plausible that parsimony$_2$ is an apseudic correlate we ask under what conditions the probability P_n of the class of n-postulate theories will

[11] A threshold objection to listing this property is that most theories do not have a uniquely determined postulate count, inasmuch as one can interchange postulates and theorems, primitives and defined terms, combine or split n-term postulates, etc. Adoption of nonoptimal conventions, while a source of "error," will *not invalidate* the count (all the indicators are imperfect) and the criterion correlation can evaluate alternative conventions. Further, in empirical sciences nonformal considerations, e.g., causal direction, Comte's Pyramid, reduction, explanation, will often direct postulate/theorem choice.

[12] I am not entirely clear about the concept of logical probability as used, e.g., by Popper, although it is obvious that the probability $P(a \,.\, b \,.\, c) \geq P(a \,.\, b \,.\, c \,.\, d)$ unless postulate d is entailed by the conjunction of the first three and so is redundant. It seems odd—I do not say inconsistent—that Popper allows speaking of a theory's 'probability' before we consider evidence bearing on it, but insists that we shift to 'corroboration'—*not* a probability—when we have evidence.

exceed the corresponding P_m of the class of m-postulate theories ($m > n$). Any such derivation of the T, F tally distribution in our finite theory table depends on the numerous latent stochastic parameters of the system *mind in society in the world* MSW[13], whose structure and dynamics are largely unknown to us. It is fruitless to approach this parameter problem via a quantitative theory of that interaction.[14] All we can presently do is derive broad (weak) inequality conditions for the parsimony$_2$ conjecture and then ask how plausible violations of these conditions appear to be.

The simplest case is that in which the base rate of *single* true postulates in a theory array is the same for theories of different "sizes" [n = number of postulates] and does not depend on how many other postulates in the theory row are T. Then the conditional truth frequency of postulates in row position k preceded by $(k - 1)$ slots filled 'T' does not differ from that for postulates preceded by, say, $(k - 3)$ true postulates. The apseudic probability of theories of size n is simply the product of the conditional probabilities

$$P_n = p_1 \times p_{2.1} \times p_{3.2} \times \ldots \times p_{n(n-1)} = p_1^n$$

a monotone decreasing function of n. If the conditional probabilities fall as we move through rows, the desired decline of P_n with n holds *a fortiori*.

An adverse scenario (counter parsimony$_2$) could arise if the conditional probability of a $(k + 1)^{th}$ postulate, in a theory of size $k > 1$, given k Ts preceding in a theory row, increases as we move through the rows. For whatever reason, a postulate becomes more probable the more true postulates there are, so theories of larger size are more likely to be apseudic than are smaller ones. It is hard to conceive why this should be the case, but we must examine the possible parameters.[15]

If such a counter-parsimony trend were to exist, it would presumably follow a more or less orderly pattern, although not necessarily monotone. Consider a pretty far-fetched case, in which each successive conditional probability is "boosted" by a multiplier $b > 1$ as we go through rows. Let p_1 = probability (over

[13] I use MSW to denote this complex system and keep us mindful that our enterprise is empirical, concerning scientists, who function in a community of scientists that is itself part of the larger social group of humankind and who are trying to figure out what's in the world, including living organisms (such as ourselves), and how the whole thing works.

[14] But the table's *manifest* parameters, however complexly caused by deep and hidden interactions, are in principle estimable at some stage of the cliometric program.

[15] Anecdotal impressions and scientific "common sense" suggest the opposite, inasmuch as the weaker and less developed sciences require many more postulates than, say, astronomy or chemistry, and no psychologist or sociologist thinks these postulates more probably true than those of the physical sciences. This is the sort of question to which simple cliometrics can provide a sufficiently accurate answer—one is confident that it will be in the parsimony$_2$ direction.

the table of theories, all rows and sizes) that *at least one* postulate is true. The conditional probability $p[(k+1)/k]$ is the probability that if k postulates are true and there is a $(k+1)^{\text{th}}$ postulate, it is true. Each of these stepwise conditionals is inflated by the multiplier b, so the ordered conditional probabilities for an n-postulate theory are

$$p_1, p_1b, p_1b^2 \ldots p_1b^{n-1}$$

Then the apseudic-probability for theories of size n is the product of these conditionals:

$$P_n = \prod^{n} p_1 b^{i-1} = p^n b^{1+2+3+\ldots+(n-1)}$$

$$= p^n b^{\frac{n(n-1)}{2}} \qquad [1]$$

where the exponent on b derives from the formula for the sum of the first m integers

$$\sum^{m} m_i = \frac{m(m+1)}{2}.$$

To evaluate this numerically, we may set an upper bound on b for a specified range of theory sizes, since none of the conditional probabilities can exceed 1. We have, for theories ranging $n = 1$ to $n = 25$, a requirement on the last conditional in the row

$$p_1 b^{n-1} \leq 1. \qquad [2]$$

(We could safely write the inequality as '<', because unless the nth postulate is redundant, its conditional on the conjunction of the others cannot = 1.) *Numerical example*: Suppose there is an even chance that, for any theory in the table, it has at least one true postulate. Then the constraint on b is

$$(.50)b^{24} \leq 1$$

yielding a "lid" on the multiplier $b \leq 1.029$, from which we obtain the theory-apseudic probabilities via

$$P_n = (.50)^n (1.029)^{\frac{n(n-1)}{2}}.$$

The graph of this function is markedly decelerated monotone decreasing with the apseudic probabilities for theories of sizes $n > 10$ differing stepwise by minuscule decrements. For example, sampling values[16] for sizes $n = 3, 5, 10, 15$:

$$P_3 = .136$$
$$P_5 = .042$$
$$P_{10} = .004$$
$$P_{15} = .0006$$

[16] One has very little intuitive (or rational!) notion of what such probabilities are, but these at least do not seem bizarre.

The apseudic probability for the least parsimonious theories in the table is $P_{25} = .0006$, just a mite larger than Buffon's famous 10^{-4}.

Despite the large exponents appearing on b as theory sizes increase, the inequality bound on the booster prevents a turnaround in this situation, and hence the rank-difference correlation (Spearman r_S) between parsimony$_2$ and apseudic probability is perfect. The metrical Pearson $r = .59$, but it is of course a poor (underestimative) descriptive statistic here due to curvilinearity. The $(n \rightarrow P_n)$ function being single-valued, an η is not computable.

Another adverse case would have the base-probabilities p_1 go up by increments $\Delta p_1 = .01$, so that starting with $p_{1(1)} = .50$ for single-postulate theories ("conjecturing only one relation, you have an even chance of being correct"), the base rises to $p_{1(25)} = .74$ for 25-postulate theories. Assume further that the conditional probabilities are boosted by a multiplier b_n chosen as the maximum possible (constrained by $p_{n/(n-1)} \leq 1$). We have

$$b_n = p_{1(n)} = .50 + (n - 1)(.01) \quad [3]$$

$$\frac{1}{\sqrt[n-1]{p_{1(n)}}} \quad [4]$$

Using these pairs in Equation [1] the $P_{(n)}$ apseudic function is monotone decreasing, sharply decelerated, yielding a Pearson correlation of –.78, again underestimating due to the nonlinearity. The rank-correlation between parsimony$_2$ and apseudic probability is perfect. *Example*: for $n = 7$, $p_{1(n)} = .56$, booster $b = 1.10$, the apseudic probability $P_{(7)} = .13$.

Less extreme booster functions are more realistic and I have experimented with several that would be considered plausible candidates by applied mathematicians in the life sciences—e.g., a booster that falls off linearly as we move through the rows, a logarithmic function of theory size, a power function with exponent $k < 1$. Of course, all of these have a lesser tendency than the preceding cases to disfavor parsimony$_2$ as an apseudic correlate. In some of them, for example, despite the nonlinearity, one finds Pearson correlations in the middle or high .90s. There is little point in contemplating numerous possible complicated functions in our present state of cliometric ignorance, but consideration of the highly implausible and strongly adverse booster instance above warrants including parsimony$_2$ in our candidate predictor list.

As regards monotonicity of the booster function, even that cannot be confidently asserted when one is contemplating the gigantic table of all scientific theories as the collection without knowing function forms, let alone parameters, of the interacting causal factors that determine the input-output relations of the system MSW. One can imagine states and processes that would lead to aperiodic "ups" and "downs" in a functional parameter like b, or one playing a comparable role in some other inflationary influence working counter to the desired parsi-

mony correlations. An obvious factor here would be that different sciences intrinsically involve very large differences in typical theory sizes, e.g., behavior genetics vs Newtonian astronomy. The various disciplines that differ systemically in this respect are also at different stages of scientific development. Such interacting factors—sometimes pushing in the same direction and sometimes oppositely—in different historical periods could generate a rather complicated wave form (such as we are accustomed to seeing in stock prices, evoked brain-cortical potentials, meteorological measures, or the oscillograph of a saxophone's tone). This possibility does not appreciably weaken the pro-parsimony derivations above. Imagining one or more such imperfectly correlated wave forms obtained by plotting inflationary parameters against theory size, since the apseudic probabilities involve multiplication, the deviations of such a wave pattern from a linear secular trend will tend to *reduce* the booster effect (for the same reason that the geometric mean is always smaller than the arithmetic).

In case the reader worries that these lines of argument prove too much, I hasten to reassure that I am not attempting to derive fallaciously a factual claim from the tautologies of the probability calculus. It is logically possible that the system MSW may be such that parsimony_2 is not an indicator of verisimilitude, either more broadly or in the narrow sense of apseudicity being considered here. Defying Einstein's dictum about the Lord God not being malicious, suppose the Big Crazy Committee in the Sky is bent on fooling us.[17] Suppose that the Committee has a special fondness for the number 7 (seven sacraments, seven wonders of the ancient world, seven virtues, the seven deadly sins, the lucky seven in throwing dice). Suppose that they have hard-wired the human brain so that in a particular scientific domain nobody invents theories with more than seven postulates. Imagine that for theories of size 1–6 postulates the conditional probabilities are constant, as in our ideal simple case with $p_1 = .50$, but given the Committee's cathexis of the magic number 7, the conditional probabilities for postulates in theories of that size are all equal to 1. Imagine further that the distribution of theory sizes runs 3–7 and is approximately normal (discretely approximated by the symmetrical binomial). In such a crazy situation, the correlation between theory size and apseudic probability is $r = .42$, a substantial negative relationship between parsimony and apseudicity. If the distribution of theory sizes were rectangular instead of quasi-normal, that unparsimonious correlation rises to .64. If the conditional probability of seven-postulate theories were reduced from $p = 1$ to $p = .90$, the rectangular distribution correlation is still positive in size, having fallen to .54 and the quasi-normal $r = .30$. But it is reassuring to note that even in this counterepistemic wiring by the malignant pantheon, if the conditional prob-

[17] Consider, for instance, Descartes' imaginary demon, or William Jennings Bryan's argument in the Monkey Trial that God put the dinosaur bones in the rocks to test our faith.

ability for postulates in a 7-postulate theory is set at $p = .75$, so that the apseudic probability of the 7-postulate theories falls to .133, then the correlations between size and apseudic probability do acquire a negative sign. Finally, if the conditional probability for the seven-postulate theory is set at $p = .65$, then we obtain high negative correlations for both the rectangular ($r = -.74$) and quasi-normal ($r = -.77$) size distributions. The amount of "rigging" required to countervail parsimony_2 is reassuring.

Parsimony_3: *Economy of theoretical concepts*.—When scientists invoke parsimony in theory appraisal, it is not always clear whether they mean having fewer law-like statements (postulates here) or having fewer theoretical concepts that appear in those statements, or both. But the "head count" is obviously not the same, and how many distinguishable theoretical concepts are present in a set of theories of equal postulate number may vary widely. There is some sort of structural (formal and semantic) feature of theories involved here, a kind of "interknittedness" or "concept density," that depends upon the amount of overlap in concepts between different postulates and the mean and dispersion of postulates' pervasities (in how many of the derivation chains that terminate in operational *wff*s[18]—*well-formed formulas*—a postulate appears essentially). It would be pointless here, pending preliminary studies in the cliometric program, to speculate about the correlation between number of postulates and number of concepts over the range of theories. We can be confident that the concept:postulate ratio will differ over research domains. That the two kinds of theoretical economy will be imperfectly correlated is obvious from the armchair, or by contemplating a sample of scientific theories, without doing a cliometric literature sample.

Some correlation must exist because there are admissibility constraints on the ratio. If there is a very large number of concepts in relation to the number of postulates (the density or interknittedness is extremely low), the theory will be marginally admissible because it has become too much like a postulate "heap" and not enough of a network. Empirical fecundity, as a matter of formal logic and mathematics, requires *overlap* between postulates in order for anything to be derived. For example, consider a k postulate theory in which each postulate is biconceptual; if there were $2k$ concepts, there would be no statement overlap and hence no derivations. In more complicated cases, the distribution of concepts over subsets of the postulates could be concentrated in such a way that some

[18] "A well-formed formula is called a *wff* by the logician and is pronounced rather like a dog barking, 'woof!' A *wff* is simply a statement that does not violate formation rules of the language being used. It may be either true or false, and (whichever it is objectively) we may or may not be in a position to decide which" (Meehl, 1992a, fn 17, p. 379). An *operational wff* is a well-formed formula in the observation language, composed of terms of the ordinary physical thing-language (Carnap) and the scientific instrument language. This use of 'operational' does not involve 'operationalism' as a philosophy of science (which I reject).

postulates are forced by sheer combinatorics to be semantic isolates. At the other extreme, extremely high conceptual densification, we run into an opposite danger because pairs, triads, etc., of concepts cannot be freely assigned to various postulates relating them without generating contradictions. *Example*: If postulate P_{12} relates quantified concepts θ_1 and θ_2 by a function $\theta_2 = f(\theta_1)$, and postulate P_{23} relates concepts θ_2 and θ_3 by $\theta_3 = g(\theta_2)$, and these relations are nomological, then we cannot freely invent a postulate P_{13} relating θ_1 and θ_3 because their relation will flow as a consequence of the first two postulates. Even if the quantitative relations postulated are stochastic rather than nomological, there are still constraints despite the "play" found in multiple correlation theory. *Example*: If the correlation of each of two variables with a third is equal, say, $r_{13} = r_{23} > .707$, then $r_{12} > 0$.[19] Investigation of the logical and mathematical relations involved in an interknittedness index—which, so far as I know, has not been looked into by either logicians or mathematicians—would be a worthwhile metatheoretical venture. For some possibly fruitful inner-structural features that might be verisimilitude correlates, see Meehl (1992a, p.379–380).

The plausibility argument for expecting concept economy to be an apseudic correlate is identical with that for parsimony$_2$. The conjunction of 'there are electrons,' 'there are protons,' 'there are positrons,' 'there are…' is generally riskier when there are more conjuncts. As in parsimony$_2$, the more you say, the more dangerous it gets. All of the discussion about boosters, bounds, trends, etc., concerning parsimony$_2$ applies *mutatis mutandis* to conceptual economy. One might suppose that a simple count (or even a statistically standardized count) of concepts would be meaningless or predictively useless unless normed in some sort of numerical relation to the number of operational *wff*s derived from them. But here again, we do not complicate matters thus at this stage, because we intend this to be taken care of by the statistics. Number of facts derived will appear later in our candidate list of apseudic indicators. Following the principle of starting simply, and because "you can't do everything at once," we are also mindful that if a candidate indicator is initially encapsulated in a composite involving another property or relation, the statistics will not enable us to disentangle it if the composite function chosen (nonactuarially) was unwise or, at any rate, nonoptimal; whereas if a candidate indicator interacts predictively with another (and so cannot be handled by a linear regression equation or discriminant function), that statistical refinement can be made if the cliometric results warrant it. The strategy at this stage is to decide what goes into the candidate list and to keep it measurably separate so that its indicator weight and interactions with the other retained indicators can be investigated.

---

[19] The mysterious quantity .707 is simply $\sqrt{2}/2$ arising from the algebra of multivariate correlation theory.

By 'number of concepts' we refer not, of course, to the number of particulars, but to the number of properties and relations or, if you prefer, the number of natural kinds defined by these predicates. We can explain the properties of macro objects to a large extent by reference to their parts, including micro parts such as molecules, and we do not require the number of molecules that explain the solid state properties of steel rods to be small in relation to the number of steel rods, there being many more molecules in one steel rod than there are steel rods in the world.[20]

There is an irksome technical difficulty for those who take the Ramsey sentence as the way to define theoretical terms implicitly, eliminating the usual theoretical terms without thereby eliminating the theory. The Ramsey Sentence is a technical device of logicians by which the theoretical terms are implicitly defined by their role in the network of postulates (Maxwell, 1962, 1970; Carnap, 1966; Lewis, 1970; Stegmüller, 1979; Glymour, 1980; Watkins, 1984). It is important because it shows how a system of expressions can *define* and *assert* concurrently, saying what a term means and asserting the existence of the theoretical entity the term denotes.[21] That the meaning of a term 'Ramseyfied out' is given solely by "upward seepage" (as I christened it in discussions in the early days of the Minnesota Center for Philosophy of Science) from the operational *wff*s is, in my view, doubtful. I divide the embedding text for a theoretical formalism into two parts: the *operational text*, in which a proper subset of the Ramseyfied theoretical terms are linked directly to observational predicates and

[20] Nolan (1997) argues from history of science examples that "quantitative parsimony," the number of individual entities of a given kind, is also desirable.

[21] While Ramsey was apparently trying to implement Russell's methodological dictum, "prefer logical constructions to inferred entities," the entities thus implicitly defined are usually considered to be inferred entities. The dummy variables, despite their theoretical nature being defined by their role in the network, are, nevertheless, bound by the existential quantifier, hence an existence claim is being made, despite this technical device for doing the semantics. I doubt that many scientific theories have been Ramseyfied, which is, of course, not the purpose of the logician or philosopher of science in discussing it. But in a field such as psychology, it has a powerful pedagogical use in making clear in this formal way how it is possible for a set of theoretical expressions to both *define* and *assert* concurrently. Many psychologists, having been brainwashed by a simplistic kind of operationism as undergraduates—not to mention having been told in high school English that a definition cannot be an assertion (of course, perfectly true for a single sentence, but not for the conjunction of sentences constituting an interesting scientific theory)—want to parse all theoretical assertions into nominal definitions on the one side and, given such nominal definitions, empirical statements on the other. This in general cannot be done, and there is no need to do it; but students have in mind the simple case that "all crows are black" cannot be a factual claim about crows if "black bird that caws and eats carrion" is the definition of 'crow.' Explaining the Ramsey sentence intrigues, illuminates, and satisfies them.

functors, and what I call the *interpretive text*, that is not operational but that *characterizes* the theoretical entities by some fairly general, but still not meta-linguistic, predicates. These generic predicates (e.g., 'combine,' 'inner,' resist'; see Meehl, 1990b, p. 4 for a list of some three dozen) are in the scientific object language, but they cut across fact domains. Sentences using these highly generic terms aid both in interpretation of the formalism and, often, warrant steps in the formalism that cannot be taken on the basis of its transformation rules alone. If I say 'θ_1 accelerates the effect of θ_2 on θ_3,' one knows how to state that relation formally in terms of a second order mixed partial derivative of the causally influenced variable with respect to the other two. One does not know whether the subject matter is economics, Freudian libido theory, or an epistatic effect in genetics. These generic terms contribute, in my view, to the understanding of a scientific theory, even when they are not needed to justify steps taken in the formalism itself, and I am not confident that such transdomain object language terms can be Ramseyfied out without loss of meaning.

Accepting the usual view, the difficulty is that when a theoretical concept θ_1 is thus implicitly defined by its role in the net (which means via its relations with the other θs), then the conditional probability for the existence claim about θ_2, when it follows the existence statement for θ_1 in our long conjunction and θ_1 and θ_2 are linked, must be $p(\theta_1 \mid \theta_2) = 1$. If the very *meaning* of one θ is given by its relations to the other θs, then if the latter do not exist, it cannot exist either, which defeats our purpose here in justifying parsimony$_3$. If we talk the usual theoretical language of a certain science, we are not in trouble, but we get into trouble if we take the Ramsey sentence to be a completely adequate account.

I am not, for the above reasons, entirely satisfied with my solution to the definitional problem, but here it is, for what it is worth: We are going to Ramseyfy out all of the θs and consider the existence statement, "There are so-and-so's," in the scientist's usual theoretical language. We substitute the variable θ_i for "so-and-so" throughout the conjunction of postulates that constitute the theory. Scanning the postulates in which θ_i occurs, we identify the set of concepts $\{\theta_{j(i)}\}$ that are related to θ_i, not indirectly (via the network), but explicitly in single postulates.

Let θ_i and the set of concepts directly related to it be $\{\theta_{ij}\}$. Then we form the Ramsey sentence of the postulates relating the $\theta_{j(i)}$ to one another, not including the postulates containing θ_i itself. This partial Ramsey sentence is $R\{\theta_{j(i)}\}$. Then, if $R\{\theta_i\}$ is the partial Ramsey sentence of the postulates containing θ_i , we consider the open formula, $R'\{\theta_i\}$, which is $R\{\theta_i\}$ with the existential quantifier struck. Finally, we write an existence claim for θ_i as follows: $(\exists\theta_i)[R\{\theta_{j(i)}\} \supset R'\{\theta_i\}]$. This sentence asserts the existence of θ_i by saying there is a θ_i such that if some other θs exist that relate to one another in such-and-such

ways, then θ_i relates to them in so-and-so ways. I am assuming that only a proper subset of all the connections of θ_i constitute what Carnap called its 'meaning postulates,' rather than saying that *all* law-like relationships in which θ_i appears, whether postulates or theorems, constitute its 'total theoretical meaning,' as has been argued against Carnap by, e.g., Sellars (1961; and cf. Maxwell, 1961).

Parsimony$_4$: *Ockham's Razor*[22].—I am using the (apocryphal?) "*Entia non sunt multiplicanda praeter necessitatem*" to have a restricted meaning different from the first three kinds of parsimony; to wit, "Do not concoct theories to explain facts already explained by an ensconced theory." This is common time-saving scientific practice, and I believe most working scientists would defend it as a policy, despite Feyerabend's (1970/1988) interesting advocacy of maximum theoretical proliferation. He thinks that concocting alternative theories is healthy, even when the ensconced theory is not in such grave difficulties factually or conceptually as to constitute a revolutionary situation. I do not enter into the merits of Feyerabend's proposal, but I consider widespread scientific practice and common sense as warranting the usual policy sufficiently to justify including parsimony$_4$ in the candidate list of apseudic indicators. The shortest argument for expecting the apseudic frequency of theories concocted in defiance of this guideline to be low is simply that if the ensconced theory is correct, or has very high verisimilitude, incompatible theories are false.

It might be supposed that Ockham's Razor is unlikely to be a useful statistical indicator because its application at first appears dichotomous: This is represent-able by placing a parsimony$_4$ value [0, 1] in front of the cliometric composite function $F(x_1, x_2, \ldots x_{11})$ as a multiplier, so it operates as a *sine qua non*, trumping everything else. [This is analogous to one meaning of *specific etiology* in medical diagnosis (Meehl, 1977).] It seems odd, but there is nothing mathematically wrong with it. Although the stated form of Ockham's Razor makes it appear that admissibility of a new theory would be a dichotomous decision determined by whether an already available theory is ensconced, on closer inspection it seems to be a matter of degree. Ensconced theories, the substantial correctness of which hardly any scientists seriously doubt, almost always have a few unsolved "puzzles." Furthermore, in rare cases of firmly ensconced theories which, at a given point in time, have literally zero anomalies,[23] some will be more firmly ensconced than others, given their status with respect to the other ten indicators in our list. Standardizing the dichotomy [0, 1] with $SD = \sqrt{pq}$ over a class of

[22] This phrase is used loosely and ambiguously, especially by social scientists, sometimes to cover what I am using it to mean here, sometimes the other three kinds of parsimony, or vaguely for all four.

[23] I don't myself know of any, certainly not in psychology.

theories, we let the cliometric statistics determine this indicator's weight, thereby bypassing cliometric appraisal of the ensconced theory. Thirdly, the cliometric discriminant function score of the ensconced theory could serve as a quantification of the extent to which a proliferating theorist is defying Ockham's Razor. Of course, a maverick risk-taking scientist might "rationally" propound a new theory without subscribing to the strong form of Feyerabend's proliferation principle. I realize that what weight, *if any*, should be assigned to the "state of the competition" in appraising a theory's merits is a matter on which logicians and philosophers of science disagree, and I have the impression that there is currently a wish to steer clear of it if possible. Whatever the resolution of that dispute, it seems appropriate to include parsimony$_4$ among our candidate indicators.

Number of corroborating facts derived. Next to parsimony, I have the impression that this is the property most commonly mentioned both by scientists and philosophers, although more so by the former. Again, we idealize by dividing theories conceptually into apseudic and pseudic, intending a later refinement in terms of verisimilitude. From the huge but finite class of operational *wff*s belonging to the theory's fact domain, we first set aside those which are not derivable, nor are their contradictories, the latter constituting a pseudic theory's falsifiers. Assume that we deal only with robust, replicable, general, operational *wff*s, that is, lawlike facts in the observation language or very closely tied to it. We know that even apseudic theories are incomplete. Some true operational *wff*s derivable from T_T are not derivable from apseudic theory T, nor are their contradictories. An apseudic theory has no false operational consequences, but it is a weakness of an apseudic theory if a large proportion of operational *wff*s are not decidable on its basis. I am not going to include that *defect* as a predictor, partly because I do not know how. More importantly, we do not, except under unusual circumstances, e.g., demands of fund granting agencies, have occasion to compare two theories concerning non-overlapping fact domains. Using mammalian behavior as an example of a given fact domain, comparing the number of facts that Skinner's operant behavior theory can derive with the number that Tolman's cognitive theory can derive, or that Hull's (now defunct) learning theory can derive, is a direct measure of the size of the complementary class of undecidable operational *wff*s. Still idealizing, I am going to assume that for the robust, replicable operational *wff*s, the auxiliaries and *ceteris paribus* clause are satisfied (see Meehl, 1997b, and the next section). Of course, we anticipate a later stage of cliometric program development in which these are additional features of theory appraisal to be put into the regression equation or discriminant function.

Thus, an apseudic theory entails no false *wff*s. I assume further that all or almost all pseudic theories entail some false operational *wff*s. So long as a pseudic theory's surviving an indefinitely large body of observational tests is a rare occurrence, our plausibility argument is not invalidated because all that implies is that a highly asymmetrical statistic between pseudic and apseudic theories will be slightly attenuated.[24]

Before cliometric research we do not know, even approximately, how many pseudic theories have been proposed in various scientific domains or over all science. But we do know anecdotally, absent actuarial study, that there are thousands of them. As mentioned above, there is only one apseudic and complete theory, T_T, and many possible pseudic incomplete alternatives. This (nonactuarial) history of science fact warrants the inference that most (nearly all?) pseudic theories are detected in the long run. Many are clearly falsified in the short run—some very quickly—by one or two robust *experimenta crucis* (Popper, 1983). This is a nice example of how naturalizing epistemology can provide "good enough" empirically-based answers to epistemological questions that are hard to answer analytically. Should a logician protest that numerous pseudic theories "can survive indefinitely," despite massive sampling of the operational *wff*s, we would ask for proof of that.

We idealize research scientists as randomly sampling from the humongous class of operational *wff*s in a theory's domain and we ask, if a theory is pseudic, what is the probability of its escaping falsification? Suppose the proportion of potential falsifying *wff*s is p. Then, if we select randomly from the class of *wff*s, the probability per experiment of escaping falsification is $q = (1 - p)$ and the probability of escaping falsification in the course of n experiments is q^n. It may be objected that these outcomes are not statistically independent, since the relationship between various randomly chosen experiments, even though they are random, is something complicated arising from the internal structure of the theory, e.g., pervasity of its false postulates, conceptual interknitting. But these terrible complexities, not yet worked out analytically by logicians and mathematicians, can be set aside because their net effect on the frequency of falsified *wff*s is already fully expressed by the number q. That is, there is a proportion p of potential falsifiers, *however they are related via the theoretical network*. This leaves a residual proportion q of *wff*s that permit pseudic T to escape detection, and if we sample randomly from the whole domain, the probability of sampling n "safe" (escaping) *wff*s is q^n.

[24] I offer a theoretical proof of this in Meehl (in preparation [2002]) despite the logicians' alleged truism (whose theorem?) that an infinite number of theories can explain any set of facts. If my argument is sound, Peirce's pragmaticist definition of truth can be viewed by a scientific realist as not definitional but criterial, a near-perfect proxy.

A second objection is that we are sampling from a finite population without replacement, so that the exact expression for the probabilities of various numbers of falsifications is not given by the expansion of the binomial $(p + q)^n$ but by the terms of the hypergeometric series. Reply: The binomial and hypergeometric series do not differ appreciably for large n ("large" does not mean hundreds or thousands, but a large sample in the statistician's sense, e.g., $n > 30$). Thus, the final term (all sampled *wff*s escapers) in the hypergeometric series is negligibly different from q^n in the binomial. To worry about this approximation would be a violation of Feigl's dictum against cutting butter with a razor.

That it is reasonable for the scientist to give greater credence to thus far successful theories that have passed more tests than to those that have passed fewer can be shown in several simple ways. Contemplating diachronically a single, domain-specific and thus far successful theory that has derived no incorrect *wff*s in n trials, we continue to sample from the class of *wff*s and the theory still escapes falsification in m attempts ($m \gg n$). Since $q^m < q^n$, the probability that the theory would be still doing this well if it were pseudic has decreased. Viewing the matter achronically, employing cross-sectional rather than longitudinal statistics, and examining the track records of competing theories within a specified fact domain, the probability of an apseudic theory escaping falsification (on our idealized auxiliary conjectures) is $p = 1$ for any number of operational *wff*s tried; whereas the proportion of pseudic theories that escape falsification is a monotone decreasing function of n. A fraction whose numerator is the number of unfalsified apseudic theories in the domain (= the number of apseudic theories in it) and whose denominator is the sum of this quantity plus a quantity formed from the escape probabilities over the pseudic theories of the domain will be a monotone increasing function of n,

$$nP_{\text{apseudic}} = \frac{nN_{\text{apseudic}}(1)}{nN_{\text{apseudic}}(1) + nN_{\text{pseudic}}\bar{q}_n^n} \ . \qquad [5]$$

Normally, the scientist is appraising theories of a particular fact domain, and ideally enters the population of *wff*s by choosing a *wff* that is derivable from one theory and whose contradictory is derivable from a competitor theory. It might be supposed that although deriving a meta principle concerning a truth indicator that cuts across unrelated domains would be of interest to the philosopher of science, it is not important for anybody else. This is almost true, but not quite. Scientists may switch to different fact domains when it appears that the leading theory in an old domain is on the verge of becoming ensconced, such that further work on it will soon be less valuable, less prestigeful, and not as intellectually exciting as in

the past, whereas related domains, for which they have research competence, are now more interesting and offer a more profitable future.[25]

In pursuing naturalized epistemology, we do not ignore the social and economic factors that influence the development of science, and part of an adequate metatheoretical program is to integrate these "extrinsic" influences into the overall metatheory of scientific progress. It is obvious that fund-granting agencies take into account the apparent theoretical progress in different research domains. However informally (and unreliably, from a cliometric standpoint) such judgments of *domain merit* are arrived at, they are constantly being made. For that reason, I consider briefly a plausibility argument that, even over totally separate and unrelated scientific domains, the number of facts derived is expected to correlate with apseudicity.

Consider N theories pooled over N_D fact domains, N_{apseudic} of which are apseudic, $N_{\text{apseudic}} \gg N_D$ because there are $(2^k - 1)$ ways to form an apseudic theory by deleting postulates from a T_{T} of size k. Since $\overline{q}_n$ is the average probability of pseudic theories escaping falsification in n intradomain experiments, $N_{\text{apseudic}}\ \overline{q}_n$ is the number of pseudic theories escaping falsification. The ratio of the number of apseudic theories escaping falsification to the total number of theories escaping, since $q = 1$ for apseudics, is given in Equation [5]. For m experiments it would be

$$mP_{\text{apseudic}} = \frac{mN_{\text{apseudic}}(1)}{mN_{\text{apseudic}}(1) + mN_{\text{pseudic}}\overline{q}_m^m} \qquad [6]$$

The desired inequality, that the proportion of apseudic theories surviving m experiments is larger than that proportion for n experiments when $m > n$ is

$$mP_{\text{apseudic}} > nP_{\text{apseudic}} \qquad [7]$$

$$\frac{mN_{\text{apseudic}}}{mN_{\text{apseudic}} + mN_{\text{pseudic}}\overline{q}_m^m} > \frac{nN_{\text{apseudic}}}{nN_{\text{apseudic}} + nN_{\text{pseudic}}\overline{q}_n^n} \qquad [8]$$

[25] Even at the graduate student level, we sometimes see this kind of thing happening. I know of at least two instances in which high caliber Ph.D. candidates who had already started research on doctoral dissertations in one of the "soft" areas of psychology became disillusioned with the long-term research prospects in the area because of what they learned in my seminar and, as a result, switched faculty advisors and their career directions. In the 1950s I conducted quantitative research on psychotherapy protocols and, after tedious coding of the patients' verbal output and much complicated time series statistical analysis, e.g., type/token ratio, verb/adjective index, verb tense changes, I concluded nothing new and interesting was learned about the therapeutic process. Conjecturing that I was not clever enough, or the state of psycholinguistics was too primitive, or the available statistical procedures were inappropriate—most likely, all three!—I decided not to research this domain further. In retrospect, I know that was a wise career choice.

which, inverting and reversing the inequality, yields

$$\frac{\left(mN_{\text{pseudic}}\right)\left(nN_{\text{apseudic}}\right)}{\left(mN_{\text{apseudic}}\right)\left(nN_{\text{pseudic}}\right)} < \frac{\overline{q}_n^n}{\overline{q}_m^m} \qquad [9]$$

The left side is the pseudic:apseudic ratio for *m*-experiment subdomains times the reciprocal of that ratio for *n*-experiment subdomains, so these ratios operate to counteract each other. If the pseudic:apseudic ratio is the same for more researched and less researched domains, the left side = 1 and the inequality holds for a wide range of values for q_m and q_n, failing only when there is a pronounced bias $q_m > q_n$ in escaping detection per experiment by pseudic theories in more researched domains. That is unknown and cliometrically researchable, but I have found no plausible arguments in either direction.

The preceding inequality relates apseudic probabilities for more versus less sampled domains, a special case of a general relation between apseudic-proportion and domain research coverage. How might this look, under various parametric circumstances, broadly specified? Writing Equation [5] as the apseudic proportion in an *n-wff* domain,

$$nP_{\text{apseudic}} = \frac{P_{\text{apseudic}}}{P_{\text{apseudic}} + \left(1 - P_{\text{apseudic}}\right)\overline{q}_n^n} \quad . \qquad [10]$$

Suppose we divide the domain into, say, seven subdomains for which the single-*wff* escape probabilities are markedly different, thus:

$$q_i = \ .35 \quad .40 \quad .45 \quad .50 \quad .55 \quad .60 \quad .65$$

Then the right side is

$$Q_n = \frac{nP_{\text{apseudic}}}{nP_{\text{apseudic}} + \left(1 - p\right)\phi\left(n\right)} \qquad [11]$$

where

$$\phi\left(n\right) = \frac{1}{7}\sum^{7} q_n^n \qquad [12]$$

hence, for increasing *n* of *wff*s sampled, $\phi(n)$ is, from $n = 1$ to $n = 25$ experiments,

$$\begin{aligned}
\phi(1) &= \frac{1}{7}\ (.35 + .40 + .45 + .50 + .55 + .60 + .65) \\
\phi(2) &= \frac{1}{7}\ (.35^2 + .40^2 + .45^2 + .50^2 + .55^2 + .60^2 + .65^2) \\
&\vdots \\
\phi(25) &= \frac{1}{7}(.35^{25} + .40^{25} + .45^{25} + .50^{25} + .55^{25} + .60^{25} + .65^{25}) \ .
\end{aligned} \qquad [13]$$

Different subdomains will rarely have the same proportions of apseudic theories, and to make life difficult for our thesis, we randomly assign values of $P_{\text{aspeudic}} = .10, .20, .30, .40, .50$, having probabilities $pP_{\text{aspeudic}} = .06, .25, .36, .25, .06$ (the quasi-Gaussian symmetric binomial). The $P_{\text{aspeudic}}(n)$ graph rises steeply from .20 to .50 at around a dozen experiments, after which it bounces around irregularly depending on small fluctuations in $\phi(n)$. The correlation coefficient between number of facts derived and apseudic probability is $r = .64$, a strong but far from perfect relation, as intuition and anecdotal evidence suggest. This coefficient is considerably lowered by the random fluctuations in $P_{\text{aspeudic}}(n)$ assigned casewise by the above pP_{aspeudic} probabilities .06, .25, …, .06 . A more realistic computation would utilize a suitable smoothed value based on the probability weights, but I have not ventured on that refinement here or in the next three examples, thus the computed correlations are all likely to be underestimates.

Possibly adverse to the desired correlation are situations in which the single-*wff* escape probabilities q_i tend to rise with n (not in the diachronic sense) so that subdomains in which many experiments are done consist of theories that, when pseudic, have a higher escape probability. This is not as bizarre as it seems at first glance. Perhaps cleverer scientists working in "better" domains, e.g., genetics rather than personology, have a tendency to do more experiments per time unit, or the society provides more financial support for "clever, successful" domains than others, or few scientists bother to concoct pseudic theories in some domains—there are numerous plausible causes for such a relation.[26]

One unfavorable scenario is a fixed increment Δq in the single-*wff* escape probabilities as we move to more researched subdomains. There is a "lid" on this increment because none of the qs exceed 1. Setting the upper bound on $q_7 = .90$, i.e., in the most deceptive subdomain a pseudic theory has a 90% chance of escaping falsification by a randomly chosen *wff*, each q_i in $\phi(n)$ is boosted according to

$$q_i(n) = q_i(1) + (n - 1)\Delta q \qquad \Delta q = .01 . \qquad [14]$$

But the exponentiation in $\phi(n)$ as n rises countervails this linear growth and the $r = .67$ does not differ appreciably from that for fixed qs. A decelerated increment specified by

$$q_i(n) = q_i(1) + .25\,[1 - l^{-.18(n-1)}] \qquad [15]$$

[26] It could of course be the other way around. In psychology, a plausible—I think factually realized—causation for this is that "soft" psychology, with substantially poorer theories and feebler research tools, attracts more would-be academics than the "hard" domains because the former are intellectually less demanding, more students take such classes, so there are more jobs, Freud's dream theory is "sexier" than electrochemistry of the retina, etc.

(the parameter –.18 again determined by a q_7 lid = .90) yields $r = .79$, a higher value. The bad scenario, where the Δqs are accelerated, following

$$q_i(n) = q_i(1)\,[1 + (.0033)(n - 1)^{1.5}] \qquad [16]$$

yields $r_{q_n} = .67$ again.

I have examined numerous other setups and parameters with highly reassuring numerical results. Conclusion: Absent extremely unfavorable and, I think, unplausible assumptions about the MSW parameters, a strong correlation exists between apseudic probability and number of facts derived.

Number of discorroborating facts derived.—If Popper_1,—i.e., Lakatos' (1968, 1978) description of Popper as a "naïve falsificationist"—requiring a definitive immediate falsification *rule*, exists and is accepted, there is no statistical problem of assigning a negative weight to discorroborating facts because one such fact, admitted into the corpus, trumps all the other indicators *modus tollens*. The theory is false and is discarded. Nobody, however, thinks that scientists ordinarily do this, and I believe no philosopher (not even Popper) has advised them to make it a rule. There are, of course, instances of what Lakatos derides as "instant rationality," for instance, the immediate slaying of the Bohr-Kramers-Slater theory by the Bothe-Geiger experiment (Popper, 1983). That BKS denied a conservation law and had made no successful predictions doubtless made it susceptible to quick rejection. Whether and when such immediate strong rejection is a rational decision is a deep, hard question, and it does not easily fit my cliometric orientation. It would amount to what we call a "stop item" in the psychometrics of personality testing, where a single scored response trumps the other items collectively, e.g., a response to a single item in military draft screening that would mandate a special intensive psychiatric interview.

One defensive strategy is calling into doubt the conjunction of auxiliaries relied on in the test. A second is conceding that the theory is probably false as stated and looking to how it might be amended. Popper objected to Lakatos' "Popper_1" (discarding the theory immediately when presented with a seeming falsifier) as a caricature of his early position. He admits that "dogmatism is to some extent necessary" (Popper, 1962, p. 49; and see 1962, p. 312, fn 1). That a scientist ought to state clearly what observations would constitute a falsification does not entail a promise to discard it from any further consideration; rather we are likely to say, "It's not literally correct as it stands, but suppose …." I am certain that Popper somewhere says that a theory should not be discarded before having a chance to "prove its mettle," but I have been unable to locate the passage. It may have been in conversation during his visit to the Minnesota Center for Philosophy of Science in the 1960s. Additionally, Popper's later emphasis on verisimilitude makes an automatic immediate discard inappropriate.

We may write the corroboration formula for appraising or testing (I prefer to say 'appraising') as follows (see Meehl, 1997b):

$$T \,.\, [T_{aux} \,.\, C_p \,.\, A_I \,.\, C_n] \vdash (O_1 \supset O_2)$$

where

T : The theory of interest
T_{aux} : Auxiliary theories relied on in the particular experiment
C_p : *Ceteris paribus* clause
A_I : Instrumental auxiliaries
C_n : The particulars stated
O_1 : An observation
O_2 : Another observation

As is well known, the auxiliary conjectures[27] may sometimes be as problematic as T itself—in the life sciences often more so, but sometimes even in the exact sciences. Faced with a seeming discorroborator, a scientist who wishes (properly or not) to defend T tries to devise experiments that will zero in on the source of the difficulty, that is, to isolate which conjuncts of the auxiliary conjunction were incorrect. (For some nice examples of this process, see Mayo, 1996.) Most experiments rely, for the theoretical derivation of the expected observational result, on only a partial subset of the theory's postulates. We regularly operate in a subregion of the nomological net. So it is incorrect to say that we are always testing the whole network of our theoretical beliefs.

Recognizing that a discorroborating result does not usually trump all the other indicators, we face here the same problem we face with each of them, that is, what weight should be given to the tally of discorroborating facts?

The asymmetry between corroboration and falsification makes the first step easier than that for the tally of favorable findings. An apseudic theory, if incomeplete, leaves operational *wff*s undecided, but it generates no incorrect predictions. So the adverse tally arises from false auxiliaries in the corroboration formula. The stochastic character of negative evidence and its resulting statistical weight in our cliometric formula arises from the (unknown) proportion of auxiliary-based failures, but having to attribute an excessive number of them speaks against the theory. We do not know how to weight this theoretically, and the facts are unavailable, but the cliometric statistics do it for us.

The problematic character of the auxiliaries in biological and social sciences makes for an almost qualitative difference between their theory appraisal and that of the physicist or chemist. Except for Kuhn's "puzzle-solving" (1977), physicists usually think of experiments as quasi-*tests* of a theory, rather as

[27] I call the bracketed term the "auxiliary conjunction" or "auxiliary conjecture," not "assumption," because the latter has three meanings in statistics, and they are often conflated (Meehl & Golden, 1982; Meehl, 1992b; Meehl & Waller, 2002).

Popper says.[28] A clinical psychologist is more likely to view a single experiment as contributing to the total empirical bookkeeping, pro and con the theory. We often do not know *which* auxiliary in *which* experiments is the culprit (Meehl, 1990d).

Could we take as our discorroborating measure, to which some predictive weight is to be assigned, the *proportion* of discorroborating *wff*s? This is unsatisfactory, since that proportion is simply the complement of the proportion of corroborating facts derived. If number of corroborating facts were expressed as a proportion, an unweighted difference between it and number of discorroborating facts would amount to $2p_{\text{corrob}} - 1$, so we would have a composite index that is a linear function of the corroboration rate. In any case, this is unsatisfactory because it ignores the widely different probative weights required by the quaitative difference between *modus tollens* and the third figure of the implicative syllogism.[29] If we standardized by setting the weight on $p = 1$, and weighting $p_{\text{discorrob}} = kp_{\text{corrob}}$ where $k > 1$, our signed composite is merely $(k + 1)p_{\text{corrob}} - k$. As either of these is linear in the other, the constants must be determined empirically, as usual in psychometrics.

None of this will work in practice, because it ignores the total *mass* of evidence—how many *wff*s are examined—pro and con. So we cannot begin by representing either as a proportion, but must use the raw tally of successes and failures as the basic metric. We start with two raw tallies, n_{corrob} and $n_{\text{discorrob}}$, the correct and incorrect derived *wff*s. Here, we remind the philosopher about how psychometrics works. The *standardizing* of the metric, in both factor analysis and in the discriminant function, is done for us by the mathematics operating on the data. Both the metric and the relative weights are data based. It may be objected that this means that the weights in our predictive function will vary from one scientific domain to another, and even with the chosen width of domains. That is correct. One must get accustomed to it, as psychologists have done for a long time in fields such as personnel selection. The relative weights given to an intelligence test and a social introversion test in selection of military personnel, law school applicants, or civil servants, may vary widely, or only slightly, over various selection tasks. The psychometrician has learned to take this as a matter of course, and it is not viewed as a methodological defect.

Considering the numbers of corroborating and discorroborating facts derived (ignoring other theory attributes) also presents the daunting epistemological

---

[28] I am indebted to Professor Roger Stuewer (History of Science and Technology, University of Minnesota) for clarification in this respect.

[29] If T is theory and O is observation, the third figure of the implicative syllogism reads $T \supset O$, O, infer T—an invalid logical form. All empirical science is in this invalid figure. Hence Morris Raphael Cohen's witticism: "All logic texts are divided into two parts. In the first part, on deductive logic, the fallacies are explained. In the second part, on inductive logic, they are committed."

problem of weighing what appear to be confirming versus disconfirming experiments. Even in the physical sciences some experiments may appear to support a theory while others contradict it; in the life sciences, such empirical conflicts are the rule, not the exception. In psychology, for instance, it is often thought that a box score of, say, 7 favorable and 3 adverse findings speaks well for a theory being evaluated. While that may at times be the correct conclusion (see below), the logic of induction prevents any such easy transition.

A critic would say such a conclusion makes it too easy for the theory, that we excuse the three predictive failures on the plausible ground that some of the tests involve false auxiliaries; but in deriving the confirming facts in the seven favorable cases, we are relying on the auxiliaries being correct. Obviously, you cannot have it both ways, trusting the auxiliaries when you like the observational result and mistrusting them when you do not.

Qualitatively, that looks like a strong rebuttal; but considered quantitatively, it appears weaker. One explains away the apparent falsifications by attributing them to incorrectness in the auxiliaries, but one cannot explain with equal ease the confirming outcomes if falsity of the auxiliaries is assumed.[30] If an auxiliary is false, successful prediction of a risky observational result (within experimental tolerance) will be possible only if one or more other auxiliaries are incorrect *in a direction and by a net amount to countervail* the first incorrect auxiliary. Such a fortunate combination of errors, leading to a proper acceptance of the theory but for a wrong reason, is considerably less probable than the case of the apparent pseudo-falsifier, although one does not know even roughly how much less probable it is. Given 10 independent studies testing a psychological theory, seven of which are positive and three negative, how could one arrive at even a rough idea of how to conduct an epistemological bookkeeping job? To concoct an index on *a priori* grounds seems impossible. This kind of epistemic complexity provides a strong argument for the Faust-Meehl Thesis.

Finally, considering all the operational *wff*s in a fact domain, the derivational power of an accessible theory is the proportion p of *wff*s that it derives. Its complement ($q = 1 - p$) has two components, the proportion of incorrect *wff*s and the proportion the theory leaves undecided. Assume that the indeterminate proportion is about the same on the average for pseudic and apseudic theories, as I can think of no plausible reason for supposing otherwise. Then since the proportion of incorrect *wff*s for apseudic theories is zero, if the proportion of incorrect *wff*s for pseudic theories > 0, the correct proportions have the relation $p_{\text{apseudic}} > p_{\text{pseudic}}$. Hence, for a random sample of n *wff*s drawn from the fact domain, the expected value of $np_{\text{apseudic}} > np_{\text{pseudic}}$. Of course, this inequality of

[30] I am here assuming we deal with severe tests involving numerical point predictions or narrow interval predictions, not with the easy-to-pass, weak tests of null hypothesis significance testing (NHST).

expected values does not guarantee the inequality will hold for a given fact domain since the proportion for a sample of *wff*s is subject to sampling error with variance pq/n, so that there will be samples in which the pseudic theory appears to be doing better. The probability of this misleading relationship declines as n increases. For fixed n, considering theories ranging over different fact domains, the positive tally will be favorable to apseudic theories, how much more favorable being dependent upon the number of *wff*s sampled. Aggregating these over domains, the positive count for the whole class of apseudic theories will be greater, and relying on correlational symmetry, theories having larger success tallies will have a higher apseudic rate.

That pseudic theories will almost always have a false *wff* probability > 0 can be shown by strong theoretical argument (Meehl, in preparation [2004]). On the empirical side, pre-cliometric evidence is persuasive, if not conclusive. The great preponderance of scientific theories have turned out to be false, a conclusion we know only because they have mispredicted *wff*s. The proportion of apseudic theories over domains is the reciprocal of the average number of actually proposed theories per domain and is a subject for cliometric investigation. Absent that, we can attain a rational conviction that pseudic theories have a very strong tendency to be tripped up by the facts simply because, over considerable time, it is extremely rare for two or more theories to survive in a fact domain; that is, science does in fact almost always ensconce one theory and discard all of its competitors. If, despite my above-mentioned theoretical arguments, pseudic theories had a negligible number of misderived *wff*s and hence a good chance of long-term survival, this historical generalization would not be true.

Qualitative diversity of facts derived.—Scientists and philosophers of science give considerable weight to a theory's ability to explain and predict general observational statements in several fact domains that in a crude phenomenological sense "appear to have nothing to do with one another." I am using the phrase 'qualitative diversity' as a relational metapredicate that is also referred to in the literature by such terms as 'range,' 'scope,' 'heterogeneity,' 'variety,' and 'deductive fertility.' Although the intuition here is quite strong, it is difficult to formulate its rationale, even in ordinary scientific language. We have a commonsensical notion that, if a theory is able to derive not only many (different) facts, but a wide range of different *sorts* of facts, that power speaks strongly in its favor. The intuition is so strong and scientific practice so clear that if we could not produce plausibility arguments in justification we would conclude that something was deficient in our epistemological or statistical

thinking. Confining operational *wff*s, as I am, to statements of general form rather than the particulars that instantiate such general statements, the distinction between "different facts" and "different *kinds* of facts" is not easy to make rigorously, but, for our modest purposes, that is not necessary. A crude division can be made between sets of law-like statements (nomological or stochastological[31]) that differ in the numerical values of controlled variables but in which the observational predicates and functors are the same from one experimental context to another or from one population (patients, animal species, classes of enzymes) to another, and experiments or statistics that observe different qualities or dimensions. An animal behaviorist could select rat, monkey, pigeon as organism; T-maze turn, lever press, target peck as the operant; hunger, thirst, shock avoidance as motives. This very narrow subset of conditions already provides $3^3 = 27$ experimental contexts, even before we begin to vary quantitative features, e.g., partial reinforcement schedules. I believe that such rough bases of division are adequate for the present purpose. *Example*: Had Einstein's General Theory of Relativity (GTR) explained the anomalous perihelion of Mercury, an anomaly in the motion of the moon, and an anomaly in the path of a comet, each to a good level of numerical accuracy, this triad of derivations would not have impressed the community of physicists as much as GTR's derivation of the Mercury anomaly, the 1919 eclipse light-bending, and the red-shift. *Example*: One of the most spectacular victories for a widely disputed theory (molecules) was Perrin's famous table showing how Avogadro's number (6.02×10^{23}) "agrees" when inferred from such diverse kinds of data as the motion of a Brownian particle and the blueness of the sky (Nye, 1972; Salmon, 1984).

In the behavioral sciences, qualitative diversity has such weight with some minds that it can countervail what we normally consider powerful negative considerations, such as lack of numerical precision, severe tests, or experimental control (manipulation). I do not assert that this is methodologically wise, but merely that it is a striking fact in the sociology of knowledge. *Example*: Disciples of Freud and Skinner are about as far apart, both as to substance and method, as students of the mind could be,[32] and the different weights they give to qualitative

[31] I coined (Meehl, 1978) the neologism stochastological (analogous to nomological) as a convenient term to refer to probabilistic relations or statistical dependencies comprised of correlations, tendencies, statistical clusterings, increments of probabilities, and altered stochastic dispositions. Perhaps because clumsy, it has not gained favor; but I advocate it as precisifying.

[32] This is true despite Skinner's grudging admiration of Freud, one of the few topics in the corpus of Skinner's writings that show an inconsistency. From many hours of conversation plus somewhat surprising positive statements in his works, I confidently attest that Fred Skinner admired Freud much more than he did the behavorists of his generation (Meehl, 1992d).

diversity versus experimental quantification is a major source of their divergence. Freudians emphasize that their theory can handle the diverse facts of hysteria, obsessions, paranoia, perversions, dreams, jokes, fairy tales, folklore, art, literature, religion, tribal customs, crowd behavior, the psychohistory of political figures, and so on. Skinnerians have a powerful, precise system of behavioral laws, but the corroborative facts are almost wholly confined to the operant conditioning chamber ("Skinner box"). One can hardly persuade them to discuss maze data, and as a result they do not require themselves to explain latent learning, which is hard to obtain in the box.

The general principle is that for a nonubiquitous (incompletely pervasive) postulate, not all operational ϕs occur in *wff*s it is used to derive. Hence there are systematic, nonchance relations between arbitrary classes of *wff*s, specified by the ϕ-sets they contain, and the postulates. However the ϕ-sets defining such *wff*-classes are specified, a false postulate P_1 does not occur equally often among the derivation chains to each of them. To say more than this requires "making cases," as in probabilifying gambling odds. But a simple situation makes the point obvious. Suppose an operational predicate ϕ_1 depends on the false postulate P_1 in the strong sense that all *wff*s containing ϕ_1 derive from P_1 and are empirically false for that reason.[33] Assume there are 20 operational predicates $\{\phi_i\}$ and that all the operational *wff*s are 2-predicate law-like relations. Then there are $\binom{20}{2} = 190$ *wff*s, 19 of which can detect the falsity of *T*. One reasonable interpretation of lowered diversity is to exclude some ϕs from the class of experiments performed. Suppose we exclude three ϕs, choosing randomly[34] from the 20 available. There are 1140 ways to pick three ϕs, and 171 ways to pick the other two ϕs if ϕ_1 is picked. Hence the probability of picking a set of experiments $[\phi_1\ \phi_i\ \phi_j]$ that exclude falsity-detector ϕ_1 is $p = \frac{171}{1140} = 15$, the probability of failing to detect *T*'s falsity. If the other case, probability $q = .85$ of detecting falsity, is realized, the probability of escaping detection is a monotone decreasing function of the number *n* of experiments performed. (This is not quite q^n because we are sampling without replacement from an urn not huge in relation to the sample size, so the hypergeometric series would be in order.) The escape-probability approaches zero as *n* increases, *and not asymptotically*, as it reaches zero at some point in the experimental series, when we run out of ϕ_1-free *wff*s to try. There

[33] I neglect undoings, where another false postulate P_2 nicely "corrects" for the defective P_1 in each derivation chain, but allowing for them would merely reduce falsification probability from $p = 1.00$ to a somewhat smaller value, not invalidating the argument.

[34] This would not require that the scientist *does* it "randomly," as by a random number table or computer randomizing algorithm; it suffices that *however* the scientist does it, the selection procedure is uncorrelated with the truth of the postulates. For present purposes a small correlation—quasi-random selection—is no problem.

are 17 ϕs left (including ϕ_1) when three ϕs are excluded. This gives us $\binom{17}{2} = 136$ *wff*s, of which 136 – 17 = 119 do not involve ϕ_1. By bad luck one could perform 119 experiments without detecting falsehood, but the 120th would detect it for sure, as would all the remaining *wff*s. Over the class of false theories, the escape-probability is less than 1 and goes to 0 with increasing *n*, so the escape-probability is less for random *wff* choice than for this less diverse selection. By our symmetry principle, the truth frequency for theories less diversely tested will be lower, how much lower depending on the number of experiments performed.

Having spent more hours than I like to recall attempting to derive a completely general metatheoretical principle that theories that can derive a wide range of qualitatively diverse facts are superior, which in turn suggests that, *ceteris paribus*, a scientist engaged in theory appraisal would be well-advised to sample the observational fact domain so as to get maximal (or very high) qualitative diversity (however we index that complicated concept), I have concluded that it cannot be done, and for a very good reason; namely, that it is not so! It is easy to see why it is not so by considering a situation common in the history of science.

Considering the finite set of operational predicates in a theory's domain, one can identify subsets of operational *wff*s on the basis of which subsets of predicates the *wff*s contain. There will be a statistical relationship, however loose, between such subsets of operational *wff*s and subsets of the theory's postulates. Only a small proportion of a theory's postulates are ubiquitous, i.e., have universal pervasity, appearing essentially in every derivation chain terminating in an operational *wff*. There are, of course, a few such. *Example*: In Hull's learning theory (1943), the postulate concerning increase in habit strength as a result of reinforcement is implicitly involved in studying *any* aspect of learned behavior, such as the potentiating effect of hunger drive on response strength. If a psychologist reported a study of that latter relationship and did not report that the experimental group (hungry) differed from the controls (satiated) in having received five times as many reinforcements, this would not merely be unscholarly, it would be a violation of scientific ethics.

In the life sciences, there are few such ubiquitous postulates, and most theories have postulates varying widely in pervasity. An important personal characteristic in which scientists differ is cleverness, whether rationalized or purely intuitive, in ferreting out a theory's weaknesses and devising experimental arrangements to bring them to light. Suppose a highly insightful scientist has a strong subjective conviction that Postulate 1 in a 10-postulate theory is erroneous but is inclined to believe that the other nine are all correct, although this postulate set is incomplete. Thus, the researcher conjectures that the theory is apseudic except for P_1. With that conviction, and being a risk-taker by temperament or studied research policy, this researcher would not enter the vast domain of

operational *wff*s randomly, or by high diversity, imitating what most fellow scientists were up to, but would instead focus attention upon that subset of *wff*s that flow from a conjunction of P_1 with others. Our scientist would do this even if P_1 does not have high pervasity and would concentrate experimental work on a narrow observational domain. This example suffices to show why a general statement, "Always seek qualitative diversity in attempting empirical appraisals of a theory," cannot be defended. *Analogy*: In a completely "blind" search for a crashed airplane in a $100 \times 100 = 10{,}000$ square mile area, it would be foolish to confine the search to the $10 \times 10 = 100$ square miles in the northwest corner, despite the fact that low, slow flying and high density coverage would improve the chance of spotting the wreck *when it is where you are looking*. But if the last radio message strongly hinted at a special sort of terrain (waters, hills, vegetation), some concentration on regions having that terrain would be rational.

What if the scientist has no such leaning against any of the individual postulates? Then whether it would be sensible to plan a research program characterized by high diversity turns out to hinge upon the kind of diversity index one concocts. It might, for instance, be sensible to scan the collection of operational *wff*s and pick them so that *every* postulate is connected with a *wff* that we will test. Thinking of a long-term program of 30 or 40 experiments devoted to this 10-postulate theory, we might then diversify *within* subdomains in such a way that Postulate 1, which is highly pervasive, is associated with as many different other postulates as possible, a nice little problem in combinatorics. So far as I know, no logician has investigated this sort of question from the standpoint of optimizing the probability of detecting pseudic theories, and it is pretty sure to be a terribly difficult logical problem.[35]

Idealizing so as to consider the *ceteris paribus* clause and the auxiliaries unproblematic, and assuming negligibly few derivational undoings of the kind that tend to make a theory inadmissible, can we say anything about a *random* entry into the collection of operational *wff*s? Not much, but a little. Because of the statistical relation that is certain to exist between relatively "homogeneous" subsets of operational predicates and the postulates that enter derivation chains terminating in them, a policy of *excluding* any considerable number of such homogeneous subsets would be unwise. Excluding k out of N such subdomains by choosing them randomly, each having selection probability $\frac{1}{k}$, or selecting excludable subdomains on the basis of the total number of *wff*s in each, will both yield a larger probability of a pseudic theory escaping detection than will a random entry into the whole pot of *wff*s. This is an algebraic truth not involving cleverness or folly, but only the scientist's good or bad luck.

[35] This seems related to Keynes' inductive strategy (1921) of increasing the negative analogy, but I have not worked it out.

The basic idea is clearly seen in the limiting case of a nearly apseudic theory containing only one false postulate P_i, such that all *wff*s containing operational predicate ϕ are incorrect and all *wff*s free of ϕ are correct. Let the proportion of ϕ-*wff*s in the fact domain be w_ϕ, so the proportion of *wff*s failing to detect pseudicity is $(1 - w_\phi)$. The number n of experiments is not involved because any member of the ϕ-set suffices to refute, and no number n sampled from the ~ϕ-set can refute. How is the probability of wrongly escaping pseudic detection related to various ways of sampling the fact domain?

The worst scenario is sampling only from the ~ϕ-set, as may be done by "defensive" scientists strongly pro-theory who (consciously or not) perceive the danger of the ϕ-set to their beloved theory.[36] The best scenario is that of a clever Popperian critic who wisely focuses on ϕ. Between these bad and good extremes lie other *wff* selections that yield different probabilities of pseudic escape. If we flipped a coin to choose between the two *wff* sets, the escape probability = $\frac{1}{2}$. Sampling *wff*s from one set only but choosing that set "quasi-randomly" in proportion to the *wff* frequencies, the escape probability is $(1 - w_\phi)$. Whether this is better than the preceding method depends on the pervasiveness of P_i. We could enter the pot of *wff*s randomly, and now the number n of experiments kicks in, the escape probability being $(1 - w_\phi)^n$. The order of escape probabilities from worst (high) to best (low) reads

$$\underset{\text{(defensive)}}{\text{avoid } \phi} > \frac{1}{2} \; ? \;\; (1 - w_\phi) > (1 - w_\phi)^n \; > \; \underset{\text{(Popperian)}}{\text{avoid } \sim\phi} \qquad [17]$$

Generalizing to a more realistic scenario, consider numerous subsets of *wff*s, defined (however!) by which combinations of ϕs they contain, that are associated with different single-*wff* escape probabilities q_i. Then the important inequality (deleting the defenders and Popperians, whose success depends on their cleverness) is

$$(w_1 q_1^n + w_2 q_2^n + \ldots + w_k q_k^n) < (w_1 q_1 + w_2 q_2 + \ldots + w_k q_k)^n \qquad [18]$$

which holds for any values of the ws and qs that lie in the probability interval [0, 1]. I do not intuit which side of that inequality should be labeled more

[36] Disciples of B. F. Skinner have valid reasons for preferring the Skinner box as an experimental instrument. But they receive a defensive side-benefit of that rational preference, inasmuch as the phenomenon of latent learning (a fairly robust effect in appropriate T-maze experiments) is hard to attain in the box; and Skinner's system cannot explain latent learning.

"diverse," but it doesn't matter so long as we are clear. Obviously a fairly "successful" defensive concentration is bad, and a clever Popperian concentration is good.

The upshot of these considerations is that a strong general claim cannot be made about selecting for diversity of *wff*s. But depending upon the scientist's ignorance and the scientific community's division of labor (Kitcher, 1990), a rough orientation toward "covering the fact waterfront widely" is likely to pay off sufficiently to warrant qualitative diversity being retained in our candidate list. One can play around with various crude indexes of observational diversity in terms of combinations of predicates, but having done some of that, I am inclined to think it is unlikely to be useful, unless the index can be derived from a logician's analysis of the formal relations of postulates to the collection of operational *wff*s. However, in the spirit of "exploratory data analysis," a case might be made that cliometric study of the predictive properties of various diversity indexes at the observational level should be part of the research agenda.

It might be supposed that, lacking a strong intuition or other source of belief in the falsity of a selected postulate, the scientist would always do best to select randomly from the entire class of operational *wff*s that belong to the domain of the theory's factual relevance. But this is not correct, either. Spelling out a general selection procedure for an optimal, or even strongly preferable, subset of operational *wff*s would require developments in the logical structure and statistical relations of theory appraisal, which does not exist and, some would say, never will. However, a plausibility argument can be made for one simple way of "covering the waterfront" that is superior to completely random entry into the whole gigantic class of *wff*s; and I believe it should be possible to generalize to more complex formulations. My example shows that a random entry into the total *wff* collection will be inferior to a *systematic covering* entry. Consider the simple case of operational *wff*s defined by pairwise correlations of operational predicates and functors. Suppose we have ten operational predicates, ϕ_1, $\phi_2 \ldots \phi_{10}$, each linked[37] closely to observational predicates ψ_1, $\psi_2 \ldots \psi_{10}$. Speaking qualitatively only, if we confine ourselves to pairwise relationships in the fact domain, we have $\binom{10}{2} = 45$ operational *wff*s. Suppose the scientist realistically contemplates being able to conduct no more than five experiments in the next five years on a research grant and, anyway, thinks it foolish to make firm plans beyond that. The task is to sample from the minuscule domain of 45 operational

[37] I bypass the interesting question of whether these relations are to be thought of as meaning postulates, purely stipulative operational definitions, or, if bridge laws, as theorems or postulates.

*wff*s.[38] We will enter this very restricted domain of 45 *wff*s randomly, and let us say that operational predicate ϕ_1 is, in all of its contexts, derived from erroneous postulate P_1. Assuming adequate numerical precision, correctness of the auxiliary theories, *ceteris paribus* clauses, etc. (Meehl, 1990a, 1997b), the false theory will be detected by an experiment involving any *wff* of the type $[\phi_1, \phi_j]$. Of the 45 operational *wff*s in the domain, only nine involve ϕ_1, so that if we were to pick a *wff* randomly from the set, we have a probability $q = .80$ of that single experiment yielding an erroneous "escape" by the pseudic theory, none of the other ϕs depending upon P_1, the only false postulate. The exact combinatorics gives a probability of escaping detection = .31; ignoring the sampling without replacement yields[39] $.80^5 = .33$. This danger, which is discouragingly probable for the scientist's hopeful five-year plan, can be avoided when we have as many as five nonoverlapping pairwise relations (I, of course, chose the number of experiments with this in mind) by associating ϕ_1 with another ϕ, say, ϕ_2, then associating ϕ_3 with ϕ_4, and so on, making sure that every one of the ten operational predicates is studied in at least one experiment. This effort to "cover the waterfront" extends into a larger number of experiments as well, where the combinatorics permits choices, still avoiding outright duplication of *wff*s involving the same predicate pairs.

To derive a general "coverage" expression requires further specification as to the logical structure of the theory in relation to its derived operational *wff*s, but I suspect that would be beyond my powers. However, by way of a slight extension into a larger number of experiments, consider this case: We plan to perform 10 experiments, not randomly chosen, and we lack strong intuitions or rational grounds for focusing on those *wff*s that involve a particular postulate as suspect. Suppose the real situation is that ϕ_1 is based upon P_1, which is false, but P_2, while true, is weak; that is, it has less power to assess verisimilitude so that, when combined with other postulates, it does not generate operational *wff*s that are quantitatively precise enough to constitute severe tests for a false theory. Then, $(P_1 \,.\, P_2) \rightarrow [\phi_1\, \phi_2]$ fails to detect false postulate P_1. Similarly, ϕ_3 depends upon postulate P_3, which is false; but P_4, while true, is also weak, and so the combination of $(P_3 \,.\, P_4) \rightarrow [\phi_3\, \phi_4]$ fails to detect. When we come to the second

[38] This is minuscule in comparison to the task we find in any empirical science, even when the theoreical domain is narrowly specified. A psychologist testing my theory of schizotaxia (Meehl, 1962, 1990c) is immediately considering, say, 10 neurological, 10 psychophysiological, 10 cognitive, and 10 psychometric candidate indicators, which, considering eight family correlation relations (parents, offspring, siblings, MZ twins, DZ twins, foster children, foster siblings, spouses), yields 80,000 pairwise operational wffs.

[39] Note the small difference, unimportant for these purposes, even with the unusually constricted fact domain. This should reassure readers troubled by some of my free-wheeling probability multiplications.

phase, using another five *wff*s to study, the pair (ϕ_1 ϕ_3) is likely to detect unless we have a forbidden countervailing, but the pair (ϕ_2 ϕ_4) fails to detect because P_2 and P_4 are both weak. Now, let P_7 and P_8 be *true* and *strong*. If we associate (ϕ_1 ϕ_7) and (ϕ_2 ϕ_8) or (ϕ_1 ϕ_8) and (ϕ_2 ϕ_7), both will detect. The point is that *we want to get out of the set* (ϕ_1 ϕ_2 ϕ_3 ϕ_4) defined by being four pairwise associations based on only four ϕs, whereas in the other way, we have four pairwise associations based upon six ϕs. The more we "spread the operational predicates around" in our sampling of operational *wff*s, we tap into more different postulate combinations. In terms of the nomological net, we will do better by not confining ourselves to subregions of the net in testing. It is obvious that among many intuitively plausible indexes of diversity that could be set up and are not mutually derivable, some will be more efficient at falsity detection than others. I will not pursue that index problem further here.

Without attempting a rigorous treatment of the general case, it is illuminating to consider the efficiency of various quasirandom selections of subsets of *wff*s that differ from a truly random choice of single experiments from the entire universe of *wff*s. Whatever method is used by the individual scientist in the community of scientists to select subdomains of *wff*s to study, the general formula for the single experiment escape probability on completely random entry is $\Sigma w_i q_i$, where w is the relative frequency of *wff*s of kind i, and q_i are their escape probabilities. The element of arbitrariness involved in how one slices the pie ("kinds of *wff*s" above) at the operational level in terms of the combinatorics of ϕs, while mildly irksome, is harmless here. Its nonoptimality for the epistemic aim is irrelevant in the derivations that follow because the quantity $\Sigma w_i q_i$ represents the total proportion of falsity detectors in the whole class of *wff*s. If we consider a different way of grouping *wff*s, what happens is that the qs undergo exactly such changes, so that the sum remains constant. When we slice the pie differently, we redistribute the potentially falsifying *wff*s over the subcategories and the qs undergo exactly the right alteration corresponding to that redistribution. The distinctions we are about to study are not distinctions as to some ideal distribution of *wff*s (the ideal one would, of course, be only *wff*s that can function as strong tests of the theory), but only involve what happens to the probability of escaping falsification when we select a finite number of *wff*s that constitute a sample from the whole domain. The escape probabilities are based upon the ways in which true and false postulates enter into the various derivation chains that terminate in the subclasses of *wff*s, however these latter are specified. Except for ubiquitous postulates, almost any way of defining subsets of *wff*s by the combinations of ϕs that occur in them will yield subsets for which the escape probabilities are not identical.

One quasi-random selection to which scientific practice will surely not conform exactly, but that might be roughly approximated under some

circumstance, is that in which the scientist, without having a falsification strategy in mind (because suspicion directed toward specific postulates does not exist), chooses to perform all, or almost all, experiments on *wff*s in a single subdomain, and where the *probability of selecting that subdomain* is proportional to the number of *wff*s it contains. In such a case, the escape probability is $\Sigma w_i q_i^n$, and for values of w and q, each lying in the probability interval [0, 1], this quantity is always less than $(\Sigma w_i q_i)^n$.

Without knowing the terrible causal intricacies and stochastic parameters of the social system, MSW, we can still make some interesting conjectures and set some bounds on the results of its (nearly certain) departure from a selection procedure that picks *wff*s randomly from the huge *wff* supply. Let me say a few words about the nearly certain properties of that system. The individual scientist, in choosing and selecting *wff*s for study, is influenced by a vast and heterogeneous collection of factors, many of which the scientist does not know about, and others known but about which nothing can be done, liking them or not. In addition to the intrinsic factors that partake of rationality in theory testing as a strictly cognitive enterprise, there are powerful psychological and extrinsic factors. *Examples*: A scientist prefers apparatus of one kind to another because it is less boring.[40] A researcher dislikes statistics, hence prefers experiments that are qualitative, or that show clear-cut quantitative relations without much statistical manipulation. Geographical or other physical constraints makes some kinds of experiments difficult.[41] Identification with one's advisor results in continuation in a particular experimental tradition. Or intense dislike for one's advisor leads to never wanting to do another experiment using a particular apparatus. A striking economic example is psychologist Harry Harlow's serendipitous discovery of "learning sets" because at one time the University of Wisconsin could not afford to purchase naive animals, thus Harlow had to reuse sophisticated monkeys. Perhaps an investigator is of junior status and without external funding, thus is forced to use cloud chambers because bubble chambers are too expensive. In the social sciences, the political, economic, ethical, and even religious views of the scientist may play an important role.

With regard to one's positive or negative attitude toward a particular theory, the Popperian falsifier approach will result in a biased selection, biased favorably for our epistemic purposes in the hands of clever experimenters; but experimenters who are not so clever may be in effect choosing a quasi-random set of

[40] I have done research with the Skinner box and with multiple and single unit T-mazes, and because I am fond of animals, including the white rat, I think maze research is more fun because you can watch the animal behaving.

[41] For many years, only rats and pigeons (not rhesus monkeys) were used at the Minnesota psychology department because it was believed, rightly or wrongly, that the monkey was excessively prone to tuberculosis and pneumonia in that climate.

*wff*s. On the other hand, there are investigators who are fond of the theory, who are not trying to falsify it (or "test" it dangerously), and who prefer to occupy themselves with precisifying parameters or with Kuhn's "puzzle solving." And, of course, a big factor is the wave of enthusiasm for a new instrument, or a new statistical procedure, or a new epistemic path to a theoretical entity, which may be warranted by exciting results, but in other situations (frequently in social sciences) has the character of a scientific fad.

Contemplating this mass of rational and nonrational—even sometimes irrational—influences on the selection of *wff*s for study, I offer a couple of plausibility arguments concerning diversity. Leaving aside whether the amount of diversification can be spoken of as deliberate when contemplating the selection of *wff*s by the whole community of scientists, considering again the formula for the probability of wrongly escaping falsification, $(\Sigma w_i q_i)^n$, the community of scientists is surely not distributing the proportion of *wff*s sampled exactly in proportion with the objective *w*s. Holding the subcategory specifications fixed, this amounts to some of the *w*s undergoing increments and others decrements, where the correlated *q*s, however, remain fixed, being functions of the relation of the theory to the facts rather than the scientist's selection of which *wff*s to examine. Consider the simple case of two subclasses of *wff*s, their two associated escape probabilities undergoing alteration in w_1 by an amount Δw_1, which results in a corresponding change Δw_2 in the other class, equal but opposite in sign. The net change in falsification probability is then $\Delta w(q_2 - q_1)$. If Δw_1 is positive and $q_1 > q_2$, the net change in erroneous escape is unfavorable, i.e., positive; otherwise it is negative. If only one experiment were being performed, these shifts would balance out. But we are performing *n* experiments, and the expected value, assuming the above complex of factors over the whole community of scientists is not biased in the direction of falsification or escape, is $\frac{1}{2}\Delta w(q + \Delta q)^n + \frac{1}{2}\Delta w(q - \Delta q)^n$. But this is positively biased because the n^{th} power of $(q + \Delta q)$ rises more than the n^{th} power of $(q - \Delta q)$ falls. So, if the complex of factors leading to a redistribution of the proportions of *wff*s in the more and less dangerous subsets is not biased or only slightly so, this departure from randomness proportional to the *wff*-kind frequencies increases the danger of erroneous falsification escape.

Another way of seeing this is in terms of the correlation between the *w*s and the *q*s. It is extremely unlikely that this correlation would be zero, but I have no argument as to whether it should be expected to be high or low. Whatever it is, over the total domain of operational *wff*s, the following is another line of argument: Suppose that the net effect of the scientific community's reassignment of *w*s to the *wff* subdomains is not strongly systematic, so the correlation r_{wq} remains substantially unchanged from what it would be were the domain sampled

randomly. In the formula for the correlation coefficient, $r_{wq} = \frac{\frac{1}{N}\sum w_i q_i - \bar{w}_i \bar{q}_i}{\sigma_w \sigma_q}$, the quantities $\bar{w}$, $\bar{q}$, and σ_q are fixed by the world, but the quantity σ_w is subject to change at the scientist's will, and a decrease in diversity means an increase in σ_w.[42] Then, if the scientific community concentrates [= counter-diversifies], although not in a systematic Popperian fashion, the denominator of r will increase. In order for the correlation to remain the same, as assumed, the term $\Sigma w_i q_i$ in the numerator must increase; and hence, the erroneous escape probability $(\Sigma w_i q_i)^n$ will increase.

Employing the positive probability shrinking denominator approach, consider two subdomains of a theory's fact domain, and the various theories alternative to T_T capable of deriving the facts in one or both domains. For any fact domain, it will almost invariably be the case that there are at least some theories capable of explaining one subdomain that will not handle the other one, so that only a subset of the competing pseudic or incomplete theories will explain the facts of both domains. Thus, the total number of theories capable of explaining one or both of the subdomains is $(n_1 + n_2 + n_{12})$, and this sum is the denominator, the size of the reference class of theories for a given domain. Only one theory, T_T, is totally adequate. If we now consider a large class of theoretical domains, the number of correct theories is equivalent to the number of domains, and the number of theories, all told, is $\sum^{n} (n_1 + n_2 + n_{12})$ over the n domains, N of these tallies of the denominator being the set of T_Ts. But if we attain this diversification and consider the theories that explain the facts of both subdomains in their domain, the denominator is markedly shrunken; so that for the fixed numerator, the proportion of true theories in the whole set of domains, i.e., the probability that a theory is correct, given that it explains both domains, is increased. The same argument goes through *mutatis mutandis* for the probability of apseudic theories, there being $2^k - 1$ apseudic theories per T_T, where k is the number of T_T's postulates. This inflation occurs in the denominator, and the same effect, although not at the same rate, is achieved by decreasing the denominator through diversity. I take it as obvious that this argument goes through for more than two subdomains per domain, although I have not constructed a proof of this.

One value of diversity is to afford protection against "accidental" miscorroborations of a pseudic theory due to a false auxiliary that just happens to cancel the main theory's quantitative error so as to make the experiment come out as predicted. Suppose a false conjunction of a subset of postulates in T entails an

[42] Perhaps this seems counterintuitive, but consider the extreme case where one subdomain gets all of the experiments and all of the other subdomains get zero experiments each. This generates an extremely high σ_w compared, say, with a rectangular or bell-shaped distribution of the ws.

observational *wff* that is in numerical error by $\Delta y > 0$ that exceeds the experimental tolerance, but the postulated falsifier fails due to a false auxiliary theory A_i whose negative observational bias $\Delta a < 0$ countervails Δy sufficiently to bring the expected net error $\Delta y - |\Delta a|$ within experimental tolerance. Consider the set of auxiliaries A_i, A_j, $A_k \ldots A_n$, each of which occurs essentially in derivations of n different *wff*s in qualitatively diverse experimental contexts appraising T. To escape falsification in all these contexts the systematic errors Δa_i, Δa_j, Δa_k,…, Δa_n must each be negative, and of sizes such that each of the net errors $(\Delta x + \Delta a_i)$ lie within experimental error tolerance. There is usually no reason why the auxiliaries, being logically independent of T, should be biased in one direction rather than another or that the bias sizes, when luckily directional, should be "just right" to correct Δy but not too much. Of course, for a single theory testable in very few independent contexts, the required distribution of Δas may have a nonnegligible probability of such fine-tuned accidental matching. We need not assume a normal, or even symmetrical, distribution of the Δa_is, but only that their algebraic signs are about equally (+) and (–) over a large class of unrelated theories, and their numerical values negligibly correlated with the associated Δys. (For these weak assumptions to be false, the gods would have to be malicious, contra Einstein's dictum.) Greater experimental diversity involves sampling more of the A_is and thereby decreases the chance of escaping falsification by auxiliary countervailings. Historical examples of numerically strong pseudo-corroborations (e.g., Worrall, 1982) have been invoked against realism and against the idea of scientific progress. This is a poor argument arising from the failure to think actuarially (Faust & Meehl, 1992; Meehl, 1992c). When one collects and contemplates striking single examples of such pseudo-corroborations, it seems discouraging. But if we remind ourselves that thousands of experiments are conducted involving thousands of theories, the expected number of these bad-luck oddities is in the hundreds. That the historian can find them does not refute realism, or objectivity, or progress. It does not even speak against those things, unless these concepts require certainty, which no one today would assert.

It might be thought that these plausibility arguments for diversity prove too much because the inequalities probabilifying erroneous escape of a pseudic theory from detection seem to rule out a powerful detection strategy in which the scientist selects operational *wff*s in a highly nondiverse *and* nonrandom manner, namely, those involving a particular postulate, P_1, conjectured to be false. But there is no contradiction between the preceding proofs and a rational adoption of a critical falsification strategy. The important distinction is between a broadly

falsifying orientation, on which the proofs explicitly rely, and a focused one, in which the scientist suspects, on whatever basis, that a certain postulate is wrong. This latter concerns the researcher's experimental strategy in planning what to try. The proofs refer to a different knowledge situation, *after* the *wff*s have been sampled, cleverly or not. The focused falsifier is making purportedly rational decisions about an empirical research program.[43] There cannot be a mathematical contradiction between my conclusions above and a cautious statement about a clever Popperian focuser using a concentration falsification strategy because they do not deal with the same statistical reference classes. The meta-meta-argument of the Popperian focuser is based on the following: "If I am correct in thinking Postulate P_1 fails—and I have faith that I am a clever thinker—my odds of detecting a pseudic theory are improved by focusing my experimental research program on those *wff*s that are connected with P_1." *After* such a research program has been conducted, and the theory has not been falsified, we are in a new epistemic situation and can assert four statements: (1) the theory has thus far been corroborated; (2) it has not been as strongly corroborated as it would have been if the *n* experiments had been diversified; (3) if the theory is nevertheless objectively pseudic, P_1 is not the culprit; (4) I was apparently not as clever as I thought.

I do not attempt to assign the relative role of formal analysis and cliometric statistics in comparing plausible indexes of diversity, an important case of the general problem of index numbers. It is not a frightening task to construct a nonarbitrary list of predicates, despite the well-known logician's problem in choosing a basic language as to which predicates should be taken as primitive. From the working scientist's standpoint, I think this will usually be relatively unproblematic. *Example*: Suppose the domain of a psychologist's interest is mammalian learning. Among the predicates that specify subclasses of possible experiments, we have the organism (rat, monkey, cat, human); the operant to be learned and performed (locomotion in a maze, pressing a lever, speaking a nonsense syllable); the motive-incentive employed (hunger, thirst), or a negative

[43] Popper asserted that that there was no rationale, recipe, prescription, or even a set of guideline principles for concocting good theories. Apparently he thought that not only did no such rationale presently exist, but there *could not be* any such. I am sure he was mistaken about this, and I do not understand why he thought it necessary to give such a doubtful prophetic thesis a central role in his philosophy of science. My theory of schizophrenia was conceived partly through first listing some metatheoretical desiderata (and dangers). My invention of taxometrics stemmed directly from contemplating the epistemic tension between unavoidably open concepts in psychopathology and the desirability of severe numerical tests. Computer programs have invented Dalton's and Kepler's theories when provided the data (Langley, Simon, Bradshaw, & Zytkow, 1990). I am not aware of theories involving postulated unobservable entities that have been so concocted, but no argument shows this to be impossible.

reinforcer, whose removal strengthens the operant and whose presentation suppresses but does not extinguish it (electric shock, loud noise, bright light, social disapproval); the discriminative stimulus (click, buzzer, stimulus word); and then a host of complicated parametric properties regarding the sequencing and timing (e.g, the classic old studies on massed versus distributed practice) or, in the Skinner box, the various kinds of schedules (Ferster & Skinner, 1957). I alluded above (Fn 38, p. 381) to the huge number of pairwise relationships that immediately come to mind for a psychopathologist aiming to appraise my theory of schizotaxia. While plausible diversity indices should be set in competition with each other cliometrically, I do not, of course, exclude the more judgmental approach of raters knowledgeable in a given scientific domain, where their deeper theoretical insight might lead them to treat some experiments as more homogeneous than a crude index of predicate combinations would capture.

Novelty of facts derived.—Consider two historical situations involving a theory T that derives four facts, f_1, f_2, f_3, f_4. In the first case, all four facts were known and employed in inventing the theory. In the second case, f_1, f_2, and f_3 were used to propound the theory, and then f_4 was derived and confirmed. One need not employ cliometrics in reading history of science, current books, and journals to conclude confidently that scientists have a strong tendency to consider the latter case more probative.[44] It is therefore disconcerting to find that philosophers of science have disagreed about the rational reconstruction of this preference, and some—failing to find any—have rejected it as irrational despite scientific practice. In the Victorian period, such first-rate metatheorists as Whewell, Peirce, and Mill disagreed.[45] Popper (1935/1959) considered risky predictions an essential feature of a scientific theory. Lakatos' methodology of research programs distinguishes his three kinds of *ad hoc*ery via the sequence of novel derivations (Lakatos, 1978). Carnap does not concede any epistemological difference. So, here we have a case of first-class intellects continuing to disagree for a century and a half over a basic methodological principle.[46]

---

[44] I have conducted a ministudy of academic colleagues, corroborating the above generalization (Meehl, 1992a, p. 404, Fn. 24).

[45] For a clarifying discussion of the issue, see Mayo (1996, Chap. 8, pp. 251-293), and for a history of the controversy, see references cited there.

[46] I vividly recall the first time I had the privilege of conversing with Carnap, when he countered my Popperian position by, "But, Meehl, how could the date of learning a fact affect its logical relation to a hypothesis?" Had I been clever (and less in awe of the eminent logician), I might have replied, "It can't, but how many principles of scientific method are deducible from Whitehead and Russell?" In our simple example of pure convergence contrasted with mixed convergence and prediction, the logical relationship between the facts and the theory is of course the same, i.e., deducibility. However, one who emphasizes novelty is talking not only about that formal relation but also about a *statistical* relation between the two cases and truth-frequency or verisimilitude. The first logic text I studied (Castell, 1935) flatly states, "It is this greater psychological force,

Those who emphasize derivation of novel facts as having special probative weight do not, so far as I know, say that novelty trumps all other considerations. If T_1 is somewhat ahead of T_2 in number (or proportion?) of novel facts derived, would we prefer it despite T_2 being more parsimonious, deriving facts of greater numerical precision, and deriving facts of greater qualitative diversity? Whatever Popper might say about such a preference, we can be pretty sure that very few philosophers would accept it, and I doubt that any working scientist would. Then there is the question of the metric. Do we want a raw number of novel facts or a proportion of derived facts that are novel? Would we consider a theory that derives five facts, one novel, to be better corroborated than a theory that derives ten facts, one novel; or the other way around, because of the sheer mass of facts derived?

What makes a fact novel? I shall not discuss that issue but adopt *use novelty*, meaning that the theory was invented without using the fact in its construction, whether or not the fact had been discovered, received by the scientific community, or known to the theorist (Mayo, 1991, 1996). *Example*: A single sentence in Einstein's 1905 special relativity article suggests he had heard of the Michelson-Morley null result on ether drift, but historians of science agree that this played no role in motivating or concocting the theory.

An important episode in the history of science, examined critically by Brush (1989, 1995), is the scientific community's acceptance of Einstein's GTR following the eclipse of 1919. Contrary to a widespread impression, Brush shows that more first-class physicists were impressed with GTR's explanation of the Mercury perihelion anomaly than by the light bending result. Why they gave the "old fact" greater weight is interesting and not easy to decide. An explicitly psychological argument for the greater importance of the old fact is that the Mercury anomaly had been known for a half-century, constituting a grave difficulty for Newtonian theory, and some able minds had invented hypotheses (some quite far fetched[47]) to explain it away, without success. A point that Brush does not stress, but I suspect is important, is the numerical precision of the perihelion explanation, whereas even the somewhat selective (biased?) choice of observational values by Eddington still left a deviation from correct prediction in light bending that was at what social scientists would call the "10% level of significance." Even so, that was far superior to Newton's estimate, out several standard errors. There was the further point that the light bending phenomenon was derivable within a restricted region of the theoretical network concerning

---

attaching to the argument from prediction, that makes its use so dramatic in the annals of discovery, But, as logicians, we are not interested in the dramatic value of a proof-form; only in its probative value" (p. 213). Alburey Castell was the prototype for philosopher Augustine Castle in Skinner's utopian novel, *Walden II*.

[47] One was that the exponent 2 on d in Newton's Law of Gravity should be 2.0001 instead.

gravity, rather than a direct test of the larger theoretical conception of GTR. I am not making an argument for or against the physicists' epistemic weights, nor am I disagreeing with Brush about the case history. I am emphasizing the compensatory influence of some metatheoretical properties and relations countervailing others when the total epistemic bookkeeping is done—a general point that helps justify the cliometric approach to metatheory.

It is difficult to state exactly what metatheoretical generalization one would like to prove, first theoretically and then cliometrically. One might formulate it *ceteris paribus*, but the cliometrician prefers to avoid that handy but ambiguous and debatable evasion. Better to rely on the cliometric statistics to take care of the other indicators' countervailings and nuisance correlations. For theoretical proof that use novelty should be a truth-correlate, I consider conjointly the number of facts derived m, and the number of those that are use novel, namely, $m - k$, where k is number of facts used in inventing the theory.

We consider theories dealing with a delimited fact domain, which could be narrow, e.g., hunger-motivated behavior of white rats in the Skinner box, or broad, e.g., mammalian learning.[48] We may consider admissible, accessible, entic, or contemplated theories, so long as the numbers are finite. For each theory, T_i there is a proportion p_i of all its derivable *wff*s that are correct, so that $1 - p_i = q_i$ is the proportion of falsifying *wff*s. Assume correct auxiliaries and accurate measurements so that confirmation and falsification are replicable and robust. Let n_1 = number of apseudic theories; n_2 = number of adequate pseudic theories[49] (they fit all the facts); and let n_3 = number of pseudic theories that derive some incorrect *wff*s. Then in sampling n_i *wff*s randomly from the fact domain, if k_i *wff*s were known in concocting the theory, and p_i is the proportion of derivable *wff*s that are correct, the probability of pseudic T_i escaping falsification is $p_i^{\,m_i - k_i}$.[50] For the n_3 falsifiable pseudic theories, the expected

[48] It could even be the whole class of empirical theories. Even with such extreme cliometric heterogeneity, I predict that strong and interesting metatheoretical generalizations will hold for this huge class.

[49] As stated above, I believe these to be nonexistent or very rare, despite the logicians' underdetermination thesis and their infinite class of adequate theories. But I include the term n_2 on the assumption that the logicians are correct.

[50] This minimum amount of "psychologism"—which is not always a sin in naturalized epistemology—serves only to specify the class of potential falsifiers. We do not try to practice cognitive science or enter the psyche of the theoretician. We only assume that the theorist, in using the k_i facts, did not commit derivational errors in logic or mathematics, which entails that those k_i known facts cannot function as potential falsifiers (see Meehl, 1990b, Appendix I: "The sin of psychologism: Keeping it venial"). Against Carnap's challenge, it is not merely the date of the fact "becoming known" that matters. The point is that because the theorist used it *and validly derived it*, it cannot subsequently function as a falsifier. Hard pressed by a zealous antipsychologism critic, one might squirm out by (uncandidly) avoiding mention of the theorists' epistemic state,

value of the undetected number is $\sum^{n_3} p_i^{m_i - k_i} = n_3\overline{\left(p_i^{m_i - k_i}\right)}$. The proportion of theories that escape falsification is then

$$P = \frac{n_1}{n_1 + n_2 + n_3\overline{\left(p_i^{m_i - k_i}\right)}}.$$

For fixed numbers n_1, n_2, n_3 and fixed distributions of p_i and m_i but variable k_i, if we sort into two classes of situations where every k_i of the first set $\leq k_i$ of the second set, the apseudic probability P is larger for the first set. I doubt that anything stronger than that can be proved, so all we have is a rather weak statement that *ceteris paribus*, as thus defined, it is preferable to have used fewer of the total derived facts when inventing a theory.

Another way of looking at it is that we want the distribution $(n - k)$ of used versus use novel facts to separate pseudic and apseudic theories as clearly as possible. Consider a class of theories, some of which have passed all of the m_i tests to which they have been subjected (by sampling *wff*s from their fact domains) and others not. We have a fourfold table in which one dichotomy is pseudic/apseudic and the other is mispredicting no sampled *wff*s versus failing at least one. If p_{ci} is the proportion of *wff*s derivable from T_i that are correct, and p_n is the proportion of the m_i that are use-novel, then the expected number of n_3 pseudic theories that survive all m_i tests is

$$n_3 \sum^{n_3} p_{ci}^{\,p_{ni} m_i} = n_3\overline{\left(p_{ci}^{\,p_{ni} m_i}\right)}.$$

One cell of this fourfold table is empty, since apseudic theories fail no tests. For fixed m_i, the *phi* coefficient from the resulting fourfold table will be greater if the p_n set is larger. This being true for all of the fourfold tables with different values of m, it will be true for the table that pools them for different numbers of *wff*s being sampled.[51]

I cannot offer a rigorous general proof that whenever the *average* exponent $(m_i - k_i)$ is larger over sets of pseudic theories varying widely in detection probability q_i, the detection rate will be greater despite correlation between the exponent and the *wff* falsity rate. However, I have tried several extreme cases numerically, with potential falsifier rates ranging from .10 to .90 and extremely high nuisance correlations. These trials are reassuring in that differences in the

---

saying disingenuously, "A factually correct *wff* derivable from *T* does not falsify it, so the potential falsifiers among *m wff*s cannot exceed $m_i - k_i$ predicted," hiding the knowledge reference in an impersonal 'predicted.' But why play games?

[51] Note that for entic theories the correlation between p_c and m_i over theories will tend to be positive because apseudic theories, by generating a smaller proportion of correct *wff*s, will tend to be detected and rejected earlier, so fewer experiments testing them will be performed.

exponent $(m - k)$ tend strongly to mask the differences in base detection probabilities, whatever the correlations.

Let m_i = number of *wff*s sampled from a theory's fact domain; P_i = proportion of *T*-derivable *wff*s that are correct; p_i = proportion of *m wff*s sampled that are use-novel. Assume the scientific community samples *wff*s randomly. Then the probability of a pseudic theory wrongly escaping falsification is

$$P(\text{esc}) = P_i^{p_i m_i} \quad .$$

Considering blocks of such values for m_i = 10, 20, 30, 40, 50 *wff*s tried, with values p_i = .10, .20, .30, …, .90, and P_i = .10, .20, .30, …, .90 randomly associated in 20 pairings per *m* block; then *P*(esc) correlates r = .56 with P_i, r = –.39 with p_i, and r = –.43 with $p_i m_i$, as expected. These Pearsonian correlations considerably underestimate the closeness of relationships because the graphs are distinctly curvilinear.[52] The average escape probability of course declines with more *wff*s sampled: *P*(esc) = .23 for m = 10, but *P*(esc) = .02 for m = 50 *wff*s.

Given the weak and somewhat fuzzy character of our thesis, how should novelty be represented in the cliometric discriminant function? Here again it is helpful to think more like the industrial psychologist than the logician. Lacking a rational reconstruction to provide an optimal metric, we *try* each of them. One could use sheer number $(m - k)$ of novel *wff*s successfully predicted where the total number m has already appeared separately in the discriminant function. One could get the proportion of novel predictions, successful or not, which seems a poor bet. One could get the proportion of successful predictions that are novel. All plausible ways are to be tried out, where "tried out" means not simply as a criterion predictor but with reference to its statistically optimal weight in the discriminant function, whereby its correlation with each of the other ten predictors is taken into account. It is the purpose of multivariate statistical optimizing to take account of the predictors' pairwise relations, positive or negative; the statistics inform us how much *independent* contribution novelty makes to forecasting ensconcement.

Numerical precision of derived facts.—Lord Kelvin considered quantification an essential feature of a scientific theory: "[W]hen you can measure what you are speaking about and express it in numbers you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind" (William Thomson, Lord Kelvin, 1891).[53]

---

[52] For what it's worth, the nonlinear correlation between P_i(esc) and p_i over all 100 pairs is η = .46.

[53] While this may be a trifle too strong, as a psychotherapist, psychometrician, and erstwhile experimental psychologist, I am willing to agree that, at least for psychology, Kelvin's notion is largely correct.

Sir Karl Popper was of course not the inventor of "severe tests" in meta-theory, although he perhaps deserves credit for the most explicit and vigorous formulation of its major role in the scientific method. Most potent is the combination of *prediction of a novel fact* with *numerical precision*. Earlier writers on philosophy of science (e.g., Whewell, 1847/1966; Jevons, 1874/1958) recognized its importance. In his magisterial *Principles of Science*, Jevons, without using the words 'severity' or 'numerical precision,' lays great stress upon the idea, providing a plethora of examples from various empirical sciences in which the close agreement of numerical values in qualitatively diverse experiments is taken to provide powerful support for a concept or theory (Chap. 25, pp. 551–573). Jevons' examples are not quite as impressive as Salmon's "remarkable coincidence" in the multiple epistemic paths to Avagadro's number (1984, pp. 213-227), but the reader cannot fail to get the point.

I think the practice of working scientists in this respect is quite clear, even though they might not be able to rigorize it. The question is not whether we think the theory in its substance is somehow "antecedently improbable," which we would prefer it *not* to be, whether we are Bayesians or not. One need not be a Bayesian to dislike theories that are antecedently highly improbable on background knowledge. The improbability absent theory, *not only the theory under contemplation but any theory*—including those that nobody has conceived of yet, or the whole class of admissible theories that are not entic—is not judged (roughly, "computed") with reference to the truth or falsity of theories, but with reference to the numerical range of the observational variable, or the range of observational functions $y = f(x)$, that we are contemplating. The reference class involved in this epistemic probability number is not the class of *theories*, but the *observational ranges*. The riskiness involves the tolerance interval of the predicted numerical value in ratio to the *Spielraum*, a convenient term I adopt from the 19th century philosopher and mathematician von Kries (see Meehl, 1990a, 1990d, 1997b). Thus one should speak not of the probability of this numerical value if the theory were false, but rather the antecedent probability of this numerical fact, *all theory aside*.

If I tell you that Meehl's theory of climate predicts that it will rain sometime next April, and this turns out to be the case, you will not be much impressed with my "predictive success." Nor will you be impressed if I predict more rain in April than in May, even showing three asterisks (for $p < .001$) in my t-test table! If I predict from my theory that it will rain on 7 of the 30 days in April, and it rains on exactly 7, you might perk up your ears a bit, but still you would be inclined to think of this as a "lucky coincidence." But suppose that I specify which 7 days in April it will rain and ring the bell; then you will start getting seriously interested in Meehl's meteorological conjectures. Finally, if I tell you that on April 4th it will rain 1.7 inches … and on April 9th 2.3 inches … and so forth, and get seven of these correct within reasonable tolerance, you will begin to think that Meehl's theory must have a lot going for it. You may believe that Meehl's theory of the weather, like all theories, is, when taken literally, false, since probably all theories are

false in the eyes of God, but you will at least say, to use Popper's language, that it is beginning to look as if Meehl's theory has considerable verisimilitude, that is, "truth-likeness" (Meehl, 1978, pp. 817-818).

Or suppose I propound a genetic theory of mammalian embryology in reliance on which I claim to be able to predict the lengths of neonatal elephants' trunks with an average absolute error of .8 cm. You would not know whether to be impressed with my theory unless you knew the mean and (more important) the standard deviation of baby elephant trunks. Thus, if their mean length were 3 cm and the standard deviation 1 cm, my predicttions average a 26% error and—worse—I could do just as well by simply guessing the mean each time (Meehl, 1997b, p. 414).

The chief objection to psychologists' conventional reliance on null hypothesis significance testing (NHST) in theory testing is that it lacks the risky, severe character of tests involving numerical precision. Successfully predicting that a mean difference is on the side $\bar{d} > 0$, while not devoid of probative value, is a weak test of a theory, especially in social science where everything is somewhat correlated with everything else and parameter difference $\Delta \neq 0$ universally. Having written at length on the mindless abuse of null hypothesis significance testing I shall not repeat it here (see Morrison & Henkel, 1970; Harlow, Mulaik, & Steiger, 1997). *Example*: From my neurological dominant gene theory of schizophrenia one might predict that the "normal" siblings of schizophrenic probands would manifest a somewhat (how much?) greater fine tremor on the Dunlap steadiness test than controls. Confirming this prediction provides support for my theory, but not much because a half dozen other plausible theories could explain it equally well (Meehl, 1990a). Contrast this with taxometric analysis (Meehl, 1995) of several soft neurology indicators in which the inferred base rate of schizotaxia among siblings is $P = .50$ (within sampling error), strongly corroborating the theory of schizotaxia inherited as a dominant gene of complete penetrance.

The notion of observational precision involves a further complication. Given the imperfection of instruments, the conceptual idealization leading to numerical approximation, the intractability of formalism leading to mathematical simplifications (e.g., dropping terms of a convergent infinite series representing a function), the literal untruth of the auxiliary theories and of the *ceteris paribus* clause—all these nuisance factors mean that even in the so-called exact sciences, and *a fortiori* in the life sciences, predicted observational numbers are not expected to be precise. Calculation of standard errors does not take these matters into account, and the theory's numerical "tolerance" about a point prediction involves a spread about that point prediction that cannot be identified simply with the statistician's random sampling distribution of observational errors. As a result, a "near-miss," wherein the theory is almost right but not quite (the "not quite" means something outside the range of standard error), does not typically lead to abandonment of the theory; everybody recognizes that coming

sufficiently close sometimes constitutes evidentiary support. For a further discussion of this *Spielraum* problem, and a proposed index of numerical corroboration, see Meehl (1990a, 1990d, 1997b).

Experimenters and theoreticians from the dawn of post-Gallilean science have relied on numerical precision in appraising theories. In most scientific writing its probative force is taken for granted without any metatheoretical comment being thought necessary to justify reliance upon it. We are comfortable with the commonly used expression, "…is in excellent agreement with the experimental values." However, it is not easy to give a cogent justification for including numerical precision in the criterion list. Adhering to our aim throughout, what we would like is a direct, forward-going proof showing that, "Considering a class of theories that provide numerically precise derivations of quantitative facts and a class of theories that provide derivations of quantitative facts which are less precise, the relative truth-frequency of the former will exceed that of the latter."

I mention briefly, without development, some familiar paths that are intuitively compelling, but which lack the *direct* character of the desired relative frequency statement. First, unamended Popper: You cannot conclusively prove the theory, and you should not even speak of "supporting" or "confirming" it, but merely of failing to refute it. However, Popper does consider theories more strongly *corroborated if they have escaped refutation* by severe tests, of which numerical precision is one kind.[54] A Bayesian, arguing affirmatively to larger posterior probabilities in Bayes' formula, locates severity of tests implicitly in the assumed *low* value of the probability of the evidence absent the theory (or, on alternative theories), where it functions somewhat hidden in the second term of the denominator.[55] If the severity is formulated in terms of the statistics of sampling, a detailed development is found in Mayo (1996). Wesley Salmon (1984), exemplifying with the beautiful case of Avagadro's number, simply

[54] Popper prefers his term 'corroborate' to the more usual (inductivist) terms 'confirm' or 'support,' although in ordinary language most of us would consider them synonymous. Whether in his later formulations, emphasizing verisimilitude and improvement in our theories, he relies on what Lakatos called a "whiff of inductivism," I do not discuss.

[55] There is an oddity about the usual way of putting this, not only as done by Bayesians. It is said, "The probability of the evidence *e*, if the theory is false, is small." This reads as if a small $p(e)$ in the second term of Bayes denominator is somehow derivable from the statement "T is false." But $\sim T = \sim S_1 \vee \sim S_2 \ \ldots \vee \sim S_k$, where the S's are the theory's postulates. How could one derive a small probability number for an observable fact from this disjunction of negations? It cannot be done. I prefer to say not "If T were false…" but rather, "Absent T…" or better, "Absent [any] theory…" The way to obtain this small atheoretical probability is not by reference to the content of the theory or its conceivable competitors (mostly unknown) but by reference to background knowledge of the observational *Spielraum* (Meehl, 1997b).

refers to the numerical precision of a successful derivation as being an unexpected and remarkable coincidence if the theory lacked verisimilitude. From the philosopher's standpoint, persistent disagreement on this matter, which concerns fundamental issues, is troublesome. The philosophizing scientist might look at it differently, saying, "Well, these philosophers seem unable to settle their differences at the rock bottom level, but it is reassuring to know that despite these persisting fundamental differences, they come out pretty much at the same place, saying that it is rational for me to do what I've been doing all along. That's nice to know, not that I would quit if they thought otherwise."

Rather than rely on indirection or intuitive appeal, I provide here a short, simple, noncontroversial argument that bypasses philosophical differences and provides a direct forward-pointing proof of the desired thesis. Consider a finite class of accessible theories dealing with a factual subdomain. To make it easy, imagine first that there are only two kinds of relationships between observables x and y: $y = \mathrm{a} + \mathrm{b} \log x$ and $y = \mathrm{c} + \mathrm{d}\sqrt{x}$. We are considering only theories, true or false, that derive the correct observational relations.

Let $n_{\log}$ = number of theories deriving the (correct) logarithmic relation,
n_{squ} = number of theories deriving the (erroneous) square root relation,
N = number of factual subdomains wherein the logarithmic function is correct.

Over the N subdomains, there is only one true theory per subdomain, but numerous false theories also capable of deriving the log function. In a given subdomain the number of theories deriving the logarithmic function is $n_{\log}$. Over N subdomains, the number of true theories is N, and the number of theories (true or false) correctly deriving the logarithmic function is $\overset{N}{\Sigma} n_{\log}$. Hence the proportion of true theories among those deriving the logarithmic function is $\frac{N}{\overset{N}{\Sigma} n_{\log}}$. In a subdomain, the number of theories deriving either a log or a square root is, $n_{\log} + n_{\mathrm{squ}}$, so the proportion of theories deriving either a log or a square root that are true is

$$\frac{N}{\overset{N}{\Sigma} n_{\log} + \overset{N}{\Sigma} n_{\mathrm{squ}}} < \frac{N}{\overset{N}{\Sigma} n_{\log}}.$$

"Weakening" all the theories so they no longer specify a log or square root but merely a monotone increasing decelerated function (satisfied by either log or square root), the relative frequency of true theories among such broadly successful theories is reduced. Generalizing this line of reasoning over a much larger class of monotone increasing decelerated functions, the class of theories deriving any of these is numbered n_{mid}; and $n_{\mathrm{mid}} \gg n_{\log}$. Hence we have

$$\frac{N}{\sum^{N} n_{\log}} \gg \frac{N}{\sum^{N} n_{\text{mid}}} .$$

We conclude that the probability of a theory being true when it correctly predicts a narrower, more precise function (or class of functions) relating numerical observations is higher than if it successfully predicts a broader class. Confining attention to a particular function form, the same argument holds for numerical precision of the function's parameters. The class of theories correctly predicting $y = .7 + .3 \log x$ has a higher truth probability than the larger class correctly predicting $y = a + b \log x$, the constants being adjustable.

Moving from the classes of admissible or accessible theories to entic theories, the proof would apply if scientists chose randomly from the classes. Because of the strength of the inequality, such deviations from random choice as occur empirically could hardly be sufficient to vitiate the above result. The argument suffices to warrant inclusion of the indicator for cliometric appraisal.

Reducibility, passive: *The theory as reduced*.—Consider a theory T_1, which, if not quite ensconced, is favorably appraised on the basis of a combination of the other 10 properties and relations, being corroborated by a set of facts $\{F_{1i}\}$. This is the theory whose appraisal is of interest at the moment. Suppose we notice that there exists another theory, T_2, ensconced or favorably appraised, being supported by its set of facts $\{F_{2i}\}$. In the typical case of reduction these facts are at a different level of description than the facts supporting T_1; and we come to realize that by adopting suitable explicit definitions of the theoretical terms of T_1 in the theoretical terms of T_2, the postulates of T_1 are seen to be theorems of T_2. In such a case, we speak of the theory of interest, T_1, as having reducibility to T_2. We might speak of "downward reducibility," or, since T_1 is the reduced theory and T_2 is the reducing theory, I shall speak of the theory of interest, T_1, *as undergoing a successful passive reduction*.

That such a finding increases our confidence in T_1 can be rationalized in several ways. These rationales are *prima facie* different, and I do not consider the question whether they are in some deep sense equivalent or mutually derivable, although I think they are not. First, because T_2 entails T_1, whatever credibility the reducing theory T_2 has on its evidence gives additional weight to T_1, over and above what T_1 has been receiving from deriving its domain facts $\{F_{1i}\}$. Thus, we would be entitled to put some confidence in T_1 had it been proliferated à la Feyerabend, without any facts at its own level of explanation because it flows as a consequence of T_2, which is rationally held independently.

A second way of looking at it is the strange coincidence argument. The two theories were concocted with different sets of facts in mind, and if they are not apseudic, it is a strange coincidence that they should yield this kind of conceptual fit merely by adoption of some definitions.

Third, to make a forward-going argument, considering the class of theories adequate to the fact domain $\{F_{1i}\}$, only a subset (usually rather small) would lend itself to such reduction. So that if T_T and apseudic incomplete portions of T_T are in the contemplated set, they constitute a larger proportion of all of the theories adequate to $\{F_{1i}\}$ and hence are more probable than those not thus reducible. I have formulated this in terms of complete reducibility, but the reasoning is the same for partial reducibility, although the extent is, of course, weaker.

There seems to be, alongside of most scientists' liking for reducibility, a distaste for it in some quarters (not in physics, chemistry, physiology, or the earth sciences). I do not understand this distaste. The whole history of science includes brilliant exercises in reduction as among its most spectacular achievements (Wilson, 1998). In the behavioral sciences, a distaste for reducibility (as a claim, as a prophecy, or even as a feature of the research agenda) is sometimes ideologically based. Sometimes it flows from "turf" considerations, as with the sociologist who bristles at the suggestion that some sociological concept could be derived from psychological principles of motivation and learning, or a psychoanalytic therapist's dislike for efforts to translate psychodynamics into learning theory concepts coming from the experimental laboratory, or the refusal of some behaviorists to take into account any direct evidence from neuroscience about how a behavioral principle is related to the brain. The most earth-shaking theoretical discovery of the last half of our century is the DNA—a reduction job, if there ever was one! It is silly to say that the reduction of something in physiology to principles of physical chemistry, e.g., how the kidney's glomeruli filter nitrogenous wastes,[56] must be rejected because it is incomplete. Incompleteness is no more a valid objection to the validity and importance of a reduction than it is to any other desirable property or relation of theories.

Reducibility, active: *The theory as reducer*.—The situation is analogous to that of passive reducibility, except that now the theorems of the contemplated theory, T_1, are translatable into the postulates of the reduced theory, T_2, by appropriate explicit definitions. *Example*: Freud's theory of neurosis is couched in mentalistic terms (wishes and defenses that ward off awareness and anxiety). Arch-behaviorist Skinner, due to his critical admiration for Freud (Meehl,

[56] Research in medical schools is often "clinical" (diagnosis, prognosis, therapy), which may or may not involve reductions. But so-called *basic* research, e.g., how the kidney glomeruli work, is explicitly and unashamedly reductionist. We analyze how the microstructure and physical chemistry filter nitrogenous wastes into urine. A philosopher who accused a typical M.D. or Ph.D. researcher of being "a reductionist" would be met with a puzzled frown and glassy stare, the doctor wondering "What is this interloper complaining about? *What else* would I be doing?"

1992d), translates portions of the Freudian system into operant behaviorism (Holland & Skinner, 1961).[57]

Here again, there are three lines of argument. First, when we discover this translatibility at the interface, the class of facts that were supportive of T_2 (because it entails them) are now entailed (mediately through T_2) by our contemplated theory, T_1. So, it has made a big jump in the number of facts in its favor. Second, the strange coincidence argument applies, for the same reasons as in passive reducibility. Third, getting the desired probabilities in the straightforward direction, the reasoning is as in the preceding section, the tally of theories in a reference class. The class of theories permitting such a conceptual fit by mere definitional translation at the theoretical interface (T_2 postulates = T_1 theorems) is almost always a proper subset of the set that are adequate to the facts of its domain. Consequently, if T_T and apseudic selections from the postulates of T_T are in the set, their relative frequency is larger because the denominator reference class is markedly reduced. These arguments assume that T_T or apseudic theories formed from it are capable of the reduction, which is a plausible conjecture based upon the tremendous success of reductions in the history of science.

COMBINING INDICATORS

Given plausibility arguments for including these 11 indicators in the candidate list, the cliometric program envisages several statistical operations which I have detailed elsewhere (Meehl, in preparation [2004]) and only summarize here. We consider a theory *ensconced* when the scientific literature shows that it satisfies certain conjunctive criteria of consensual acceptance. If it remains ensconced for another 50 years, we take this 50-year ensconcement as a proxy for Peircean ultimate survival; and if a theory has been discarded for 50 years we take that as a proxy for ultimate rejection (Peircean "truth" versus "falsity"). The adequacy of the 50-year proxy is to be noncircularly tested by fitting a curve (cumulative record) of reversals during the half-century or more following the 50-year ensconcement/discard criterion. The asymptotes of these curves estimate what proportion of ensconced theories will nevertheless be (surprisingly) discarded in the long run, and what proportion of apparently slain theories will be resurrected. My conjecture is that both of these proportions will be small enough to warrant using the half-century ensconcement criterion as a proxy for ultimate fate.[58] For a consistent pragmaticist (instrumentalist), this will suffice. For the scientific realist, we need further arguments for accepting

[57] I consider these translations only partly successful, but that is not the point here.

[58] I optimistically predict that the first proportion will be < .05, and the second < .01. I cannot name a theory that was revived in its original form after a 50-year discard. Prout's hypothesis that atomic weights are integral multiples of hydrogen was never discarded, and Sir William Crookes even prophesied the correct explanation (isotopes, without using the word) (Jaffe, 1976).

Peircean survival as a proxy for verisimilitude, and I have offered such (Meehl, in preparation [2004]).

Accepting the ensconcement proxy, we conduct several statistical analyses on the candidate indicators. They include a discriminant function predicting the proxy dichotomy, principal component factor analysis (deleting the proxy), taxometric analysis (deleting the proxy), and various relations of these three to a content-based index of each theory's similitude to the ensconced theory of its fact domain. We do this for scientific subdomains and for empirical science as a whole. We do not mind computing statistics on a mixture of apples and oranges because the parameters for the subclasses are obtainable by disaggregation and, of course, do not contradict the parameters found for the mixture.

Convergence of these lines of statistical evidence, interlocking with theoretical arguments concerning (1) indicator predictive power and (2) statistical chances of false theories' detection, would warrant rational belief in the conjecture that objective truth-likeness underlies the pattern of relationships. I rely here on the basic epistemological notion of "inference to the best explanation" (Harman, 1965; Lipton, 1991), realizing that the logicians have not as yet succeeded in rigorously explicating it. I am firmly convinced that it is the core of all our reasoning in psychology and other empirical sciences, as well as in history, biography, courts of law, business, and personal affairs. The best I can do by way of rough explication is that theory T (strictly or probabilistically) implies facts $f_1, f_2 \ldots f_n$, and this conjunction of facts has a low prior probability absent T. This second feature makes it a "good" explanation. A "best" explanation somehow involves combining the smallness of that factual prior with the antecedent probability of T, based on background knowledge. Whether these relations are adequately expressed by Bayes' Rule, I shall not discuss. As to the atheoretical prior on the facts, my emphasis is on the *Spielraum* provided by background knowledge (Meehl, 1990a, 1990d). The main point is that empirical metatheory is appraised in the same way as first-order scientific theories—by its ability to explain the facts about the latter. Of course, for first-order scientific theories the ordering of competing theories as to their "goodness" is in the long run to be appraised on the basis of cliometric statistics, e.g., theories' "scores" on the linear discriminant function forecasting 50-year ensconcement.

REFERENCES

Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *2n international symposium on information theory*. Budapest: Akademiai Kiado. Pp. 267-281.

Barber, B. (1961) Resistance by scientists to scientific discovery. *Science*, 134, 596-602.

Boyd, R. N. (1973) Realism, underdetermination, and a causal theory of evidence. *NOÛS*, 7, 1–12.

Brush, S. G. (1989) Prediction and theory evaluation: the case of light bending. *Science*, 246, 1124-1129.

Brush, S. G. (1995) Dynamics of theory change: the role of predictions. *PSA 1994* [Proceedings of the Philosophy of Science Association], 2, 133-145.

Carnap, R. (1966) *Philosophical foundations of physics*. New York: Basic Books.

Castell, A. (1935) *A college logic*. New York: Macmillan.

Cohen, M. R., & Nagel, E. (1934) *An introduction to logic and scientific method*. New York: Harcourt, Brace.

Copi, I. M. (1961) *Introduction to logic* (2nd ed.) New York: Macmillan.

Dauer, F. W. (1989) *Critical thinking: an introduction to reasoning*. New York: Oxford Univer. Press.

DeVito, S. (1997) A gruesome problem for the curve-fitting solution. *British Journal for the Philosophy of Science*, 48, 391-396.

Ellis, A. (1957) Outcome of employing three techniques of psychotherapy. *Journal of Clinical Psychology*, 13, 334-350.

Faust, D. (1984) *The limits of scientific reasoning*. Minneapolis, MN: Univer. of Minnesota Press.

Faust, D., & Meehl, P. E. (1992) Using scientific methods to resolve enduring questions within the history and philosophy of science: some illustrations. *Behavior Therapy*, 23, 195-211.

Faust, D., & Meehl, P. E. (2002) Using meta-scientific studies to clarify or resolve questions in the philosophy and history of science. *Philosophy of Science*, 69, S185-S196.

Feigl, H. (1929) Meaning and validity of physical theories. [Reprinted in R. S. Cohen (Ed. & Transl.), *Herbert Feigl: inquiries and provocations: selected writings 1929–1974*. Boston: D. Reidel, 1981. Pp. 116-144.] [Original publication in German]

Feigl, H. (1950) Existential hypotheses: realistic versus phenomenalistic interpretations. *Philosophy of Science*, 17, 35-62. [Reprinted in R. S. Cohen (Ed.), *Herbert Feigl: inquiries and provocations: selected writings 1929–1974*. Boston, MA: D. Reidel, 1981. Pp. 192-223.]

Ferster, C. B., & Skinner, B. F. (1957) *Schedules of reinforcement*. New York: Appleton-Century-Crofts.

Feyerabend, P. (1970/1988) Against method. In M. Radner & S. Winokur (Eds.), *Minnesota studies in the philosophy of science*. Vol. IV. *Analyses of theories and methods of physics and psychology*. Minneapolis, MN: Univer. of Minnesota Press. Pp. 17-130. [Expanded and published as a book, *Against method*. (Rev. ed.) New York: Verso.]

Forster, M. (1995) The golfer's dilemma: a reply to Kukla. *British Journal for the Philosophy of Science*, 46, 348-360.

Forster, M., & Sober, E. (1994) How to tell when simpler, more unified, or less *ad hoc* theories will provide more accurate predictions. *British Journal for the Philosophy of Science*, 45, 1-35.

Frank, P. G. (1954) The variety of reasons for the acceptance of scientific theories. *Scientific Monthly*, 79, 139-145.

Gillies, D. A. (1973) *An objective theory of probability*. London, Eng: Methuen.

Glymour, C. (1980) *Theory and evidence*. Princeton, NJ: Princeton Univer. Press.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.) (1997) *What if there were no significance tests?* Mahwah, NJ: Erlbaum.

Harman, G. (1965) The inference to the best explanation. *Philosophical Review*, 74, 88-95.

Hempel, C. G. (1966) *Philosophy of natural science*. Englewood Cliffs, NJ: Prentice-Hall.

Holland, J. G., & Skinner, B. F. (1961) *The analysis of behavior: a program for self-instruction*. New York: McGraw-Hill.

Hull, C. L. (1943) *Principles of behavior*. New York: Appleton-Century.

Jaffe, B. (1976) *Crucibles: the story of chemistry, from ancient alchemy to nuclear fission*. New York: Dover.

Jevons, W. S. (1874/1958) *The principles of science*. New York: Dover.

Keynes, J. M. (1921) *A treatise on probability*. London, Eng.: Macmillan.

Kitcher, P. (1990) The division of cognitive labor. *Journal of Philosophy*, 87, 5-22.

Kordig, C. R. (1971a) The comparability of scientific theories. *Philosophy of Science*, 38, 467-485.

Kordig, C. R. (1971b) *The justification of scientific change*. Boston, MA: D. Reidel.

Kordig, C. R. (1978) Discovery and justification. *Philosophy of Science*, 45, 110-117.

Kuhn, T. S. (1977) Objectivity, value judgment, and theory choice. In *The essential tension: selected studies in scientific tradition and change*. Chicago, IL: Univer. of Chicago Press. Pp. 320-339. [Reprinted in Brody, B. A., & Grandy, R. E. (Eds.) *Readings in the philosophy of science*. (2nd ed.) Englewood Cliffs, NJ: Prentice Hall, 1989. Pp. 356-368.]

Kukla, A. (1995) Forster and Sober on the curve-fitting problem. *British Journal for the Philosophy of Science*, 46, 248-252

Lakatos, I. (1968) Criticism and the methodology of scientific research programmes. *Proceedings of the Aristotelian Society*, 69, 149-186.

Lakatos, I. (1978) *Philosophical papers*. Vol. I. *The methodology of scientific research programmes*. (J. Worrall & G. Currie, Eds.) New York: Cambridge Univer. Press.

Langley, P., Simon, H. A., Bradshaw, G. L., & Zytkow, J. M. (1990) *Scientific discovery: computational explorations of the creative processes*. Cambridge, MA: MIT Press.

Laudan, L. (1984) *Science and values*. Berkeley, CA: Univer. of California Press.

Lewis, D. (1970) How to define theoretical terms. *Journal of Philosophy*, 67, 427-446.

Lipton, P. (1991) *Inference to the best explanation*. New York: Routledge.

Margenau, H. (1950) *The nature of physical reality*. New York: McGraw-Hill.

Maxwell, G. (1961) Meaning postulates in scientific theories. In H. Feigl & G. Maxwell (Eds.), *Current issues in the philosophy of science*. New York: Holt, Rinehart & Winston. Pp. 169-183.

Maxwell, G. (1962) The ontological status of theoretical entities. In H. Feigl & G. Maxwell (Eds.), *Minnesota studies in the philosophy of science*. Vol. III. *Scientific explanations, space, and time*. Minneapolis, MN: Univer. of Minnesota Press. Pp. 3-27.

Maxwell, G. (1970) Structural realism and the meaning of theoretical terms. In M. Radner & S. Winokur (Eds.), *Minnesota studies in the philosophy of science, vol. IV: Analyses of theories and methods of physics and psychology*. Minneapolis, MN: Univer. of Minnesota Press. Pp. 181-192.

Mayo, D. G. (1991) Novel evidence and severe tests. *Philosophy of Science*, 58, 574-585.

Mayo, D. G. (1996) *Error and the growth of experimental knowledge*. Chicago, IL: Univer. of Chicago Press.

Meehl, P. E. (1954/1996) Clinical versus statistical prediction: a theoretical analysis and a review of the evidence. Minneapolis, MN: Univer. of Minnesota Press. (Reprinted with new Preface, 1996, by Jason Aronson, Northvale, NJ.)

Meehl, P. E. (1962) Schizotaxia, schizotypy, schizophrenia. *American Psychologist*, 17, 827-838.

Meehl, P. E. (1977) Specific etiology and other forms of strong influence: some quantitative meanings. *Journal of Medicine and Philosophy*, 2, 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, 806-834.

Meehl, P. E. (1990a) Appraising and amending theories: the strategy of Lakatosian defense and two principles that warrant using it. *Psychological Inquiry*, 1, 108–141, 173–180.

Meehl, P. E. (1990b) *Corroboration and verisimilitude: against Lakatos' "sheer leap of faith"*. (Working Paper, MCPS–90–01) Minneapolis, MN: Univer. of Minnesota, Center for Philosophy of Science.

Meehl, P. E. (1990c) Toward an integrated theory of schizotaxia, schizotypy, and schizophrenia. *Journal of Personality Disorders*, 4, 1-99.

Meehl, P. E. (1990d) Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, 66, 195–244. [Also in R. E. Snow & D. Wiley (Eds.), *Improving inquiry in social science: a volume in honor of Lee J. Cronbach*. Hillsdale, NJ: Erlbaum, 1991. Pp. 13–59.]

Meehl, P. E. (1992a) Cliometric metatheory: the actuarial approach to empirical, history-based philosophy of science. *Psychological Reports*, 71, 339-467.

Meehl, P. E. (1992b) Factors and taxa, traits and types, differences of degree and differences in kind. *Journal of Personality*, 60, 117-174.

Meehl, P. E. (1992c) 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, 267-282.

Meehl, P. E. (1992d) Needs (Murray, 1938) and state-variables (Skinner, 1938). *Psychological Reports*, 70, 407-450.

Meehl, P. E. (1995) Bootstraps taxometrics: Solving the classification problem in psychopathology. *American Psychologist*, 50*,* 266-275.

Meehl, P. E. (1997a) Credentialed persons, credentialed knowledge. *Clinical Psychology: Science and Practice*, 4, 91–98.

Meehl, P. E. (1997b) The problem is epistemology, not statistics: replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* Mahwah, NJ: Erlbaum. Pp. 393-425.

Meehl, P. E. [2004] Cliometric metatheory III: Peircean consensus, verisimilitude, and the asymptotic method. *British Journal for the Philosophy of Science*, 55, 615-643.

Meehl, P. E., & Golden, R. (1982) Taxometric methods. In P. Kendall & J. Butcher (Eds.), *Handbook of research methods in clinical psychology*. New York: Wiley. Pp. 127-181.

Meehl, P. E., & Waller, N. G. (2002) The path analysis controversy: a new statistical approach to strong appraisal of verisimilitude. *Psychological Methods*, 7, 283-300.

Merck's 1899 manual. (1999) New York: Merck.

Morrison, D. E., & Henkel, R. E. (Eds.) (1970) *The significance test controversy*. Chicago, IL: Aldine.

Newton-Smith, W. H. (1981) *The rationality of science*. Boston, MA: Routledge & Kegan Paul.

Nolan, D. (1997) Quantitative parsimony. *British Journal of Philosophy*, 48, 329-343.

Nye, M. J. (1972) *Molecular reality*. London: Macdonald.

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 Univer. Press. Pp. 257-276. (Originally published in *Popular Science Monthly*, 12, 286-302)

Popper, K. R. (1935/1959) *The logic of scientific discovery*. New York: Basic Books.

Popper, K. R. (1962) *Conjectures and refutations*. New York: Basic Books.

Popper, K. R. (1983) *Postscript*. Vol. I. *Realism and the aim of science*. Totowa, NJ: Rowman & Littlefield.

Quine, W. V. O. (1969) Epistemology naturalized. In W. V. O. Quine, *Ontological relativity and other essays*. New York: Columbia Univer. Press. Pp. 69-90.

Reichenbach, H. (1938) *Experience and prediction*. Chicago, IL: Univer. of Chicago Press.

Salmon, W. C. (1984) *Scientific explanation and the causal structure of the world*. Princeton, NJ: Princeton Univer. Press.

Schaffner, K. F. (1970) Outlines of a logic of comparative theory evaluation with special attention to pre- and post-relativistic electrodynamics. In R. Stuewer (Ed.), *Minnesota studies in the philosophy of science*. Vol. V. *Historical and philosophical perspectives of science*. Minneapolis, MN: Univer. of Minnesota Press. Pp. 311-354.

Schmaus, W. (1996) The empirical character of methodological rules. *Philosophy of Science*, 63(Proceedings), S98-S106.

Sellars, W. (1961) The language of theories. In H. Feigl & G. Maxwell (Eds.), *Current issues in the philosophy of science*. New York: Holt, Rinehart & Winston. Pp. 57-77.

Shapere, D. (1977) Scientific theories and their domains. In F. Suppe (Ed.), *The structure of scientific theories*. (2nd ed.) Chicago, IL: Univer. of Illinois Press. Pp. 518-589.

Skinner, B. F. (1938) *The behavior of organisms: an experimental analysis*. New York: Appleton–Century.

Sneed, J. D. (1976) Philosophical problems in the empirical science of science: a formal approach. *Erkenntnis*, 10, 115-146.

Stegmüller, W. (1979) *The structuralist view of theories*. New York: Springer-Verlag.

Thagard, P. (1992) *Conceptual revolutions*. Princeton, NJ: Princeton Univer. Press.

Thompson, Sir William (Lord Kelvin) (1889) Electrical units of measurement. In *Popular lectures and addresses*. Vol.1. London: Macmillan.

Todhunter, I. (1865) *A history of the mathematical theory of probability*. Cambridge, Eng: Cambridge Univer. Press.

Watkins, J. W. N. (1984) *Science and scepticism*. Princeton, NJ: Princeton Univer. Press.

Whewell, W. (1847/1966) The philosophy of the inductive sciences, founded upon their history. New York: Johnson Reprint.

Wilson, E. O. (1998) *Consilience: the unity of knowledge*. New York: Knopf.

Wilson, M. (1980) The observational uniqueness of some theories. *Journal of Philosophy*, 77, 208-233

Worrall, J. (1982) The pressure of light: the strange case of the vacillating 'crucial experiment.' *Studies in History and Philosophy of Science*, 13, 133–171.

Accepted August 5, 2002,

pdf by LJY, October 2010