#154

CLIOMETRIC METATHEORY: THE ACTUARIAL APPROACH TO EMPIRICAL, HISTORY-BASED PHILOSOPHY OF SCIENCE[1]

PAUL E. MEEHL

University of Minnesota

Summary. — Metatheory is the empirical theory of scientific theorizing. Its descriptive data base is scientific practice, history of science, and the facts of human cognition and communication. "Scientific method" is a loose set of *principles* (guidelines, policies, rules of thumb, helpful hints, preferences) plus a few strict *rules*. The analytical and prescriptive functions of metatheory try to explain scientific success and failure and to justify (rationalize) the principles as conducive to science's epistemic aims, employing the findings of behavioral science, probability theory, formal logic, and armchair epistemology as explanatory tools and constructs. Because the several methodological principles are incommensurable and their relation to our epistemic goal stochastic, metatheoretical research should supplement case studies with explicitly actuarial methods, sampling episodes from history of science and subjecting them to formal psychometric treatment. Psychologists' mental habits and quantitative skills should enable us to take the lead in developing cliometric metatheory as a new discipline.

CONTENTS

---

[1]*Psychological Reports*, 1992, 71, 339-467. Monograph Supplement 1-V71. [This is the first of three publications by Meehl that were originally intended to be a book. The other two articles are: "Cliometric metatheory II: Criteria scientists use in theory appraisal and why it is rational to do so." *Psychological Reports*, 2002, *91*, 339-404; and, published posthumously, "Cliometric metatheory III: Peircean consensus, verisimilitude, and asymptotic method." *British Journal for the Philosophy of Science*, 2004, *55*, 615-643. —LJY]

When philosopher Sneed, author of a seminal metatheoretical analysis of *concepts* in classical mechanics (1979), says flatly that philosophy of science is a branch of social science (1976), a psychologist raised—as I was—on analytic philosophy in general and logical positivism in particular finds such a statement jarring. Though in one sense Sneed's assertion is not shocking but "obvious"; science is indisputably a behavior product. Equations and text in books, scientific instruments, photographs and schematic diagrams, tables of functions and physical constants are all produced by "mind in society." Even Popper's (1972) world 3 of theories and concepts—a world the metaphysical status of which is somewhat obscure—is the resultant of world 2 (human mind) interacting with world 1 (physical nature). I know of no positivist, Popperian, or language analyst who denied this near truism or found it uncomfortable. My teacher Herbert Feigl, inventor of the term ⌜logical positivism⌝ and co-author of the first English paper on the Vienna Circle position (Blumberg & Feigl, 1931), took it for granted that psychologists study the scientist thinking and experimenting, sociologists study social influences on scientists, economists and political scientists investigate how society allocates resources to the various sciences competing for support, and so on. While these matters were not interesting to the positivists, neither were the positivists threatened by them. The threat occurs when someone asserts that the *only* legitimate questions are of this socio-psychological kind, that the traditional concerns of philosophers—conceptual analysis, validity, rational reconstruction, optimizing strategy, "armchair epistemology"—are undoable, or pointless, or must be totally reformulated in terms of cognitive psychology and sociology of knowledge. Let me be clear at the start that I do not hold that view; indeed, the position of this paper is incompatible with it. If metatheory[2] is taken to be the scientific theory of scientific theorizing, I assume its tasks include not only accurate ascertainment of the empirical facts ("episodes" in history of science) and making generalizations about them, but also *explaining* scientific development. In particular, a satisfactory metatheory should explain why science "works better" than any other putatively cognitive enterprise, why post-Galilean science works better than medieval (*pace* Feyerabend), why it fails when it does, why some sciences grow better than others, and so on. I believe that no *explanation* of these things is possible if the explanatory toolkit is deprived of the apparatus of logic, mathematics, probability theory, and even "armchair epistemology."[3]

---

[2]I prefer the term ⌜metatheory⌝ to the more usual ⌜philosophy of science⌝, partly because of its grammatical convenience, lending itself to adjectival form; partly because it directly suggests that metatheory is *the (empirical) theory of theorizing*, the current view; and partly for propagandistic reasons in my own discipline, because psychologists—particularly of the "hard-nosed" (behavioristic, psychometric, or biological) orientations—tend to be suspicious of philosophers, or at least doubt that there is anything useful to learn from them.

[3]By "armchair" in this context I intend nothing derogatory. (As Bertrand Russell says, an armchair is an excellent place in which to think!) Social scientists have a tendency to assume the equivalence: empirical = experimental = quantitative (cf. Meehl, 1971) and this conflated concept is then contrasted, invidiously, with "armchair." Properly used, the term ⌜empirical⌝ means *based on experience*, and obviously most of our experiences are neither experimental or quantitative. Within the sciences, some research domains are not experimental (e.g., astronomy, geology, epidemiology) and others are only rarely and weakly quantitative (comparative anatomy, botany, anthropology). The great armchair epistemologists began their reasonings with certain well known and undisputed *facts* about the world, the human mind, and their relations. Everyone agrees that humans learn about the world (including other minds) by seeing, hearing, touching, etc.; that we can remember past experiences, but fallibly; that the same object appears different from different distances and angles; that a stick looks bent in water but feels straight to touch; that humans have

In this paper I propound and defend a strong metatheoretical thesis: *The relation of metatheory to empirical history of science is intrinsically statistical; hence our meta2-appraisal of metatheories should proceed actuarially, sampling randomly from episodes in the history of science and analyzing relationships by appropriate statistical and psychometric methods*. So far as I know, hardly any historians or philosophers of science think of the problem in this way.[4] If the idea is unsound, the process of showing why it is mistaken, will, I am sure, be illuminating on all sides. If it turns out to have merit, even if only as adjunctive to the received methods (e.g., case studies, opinion surveys, statistical studies of the *Science Citation Index*), an important corollary would be: *Within an acceptable metatheoretical frame, the appraisal of specific scientific theories should also be explicitly actuarial*.

It is now generally admitted, even by the remaining living members of the "logical empiricism" school and scientists influenced by them, that philosophy of science should

---

biases and make mistakes; that some external event sequences are highly predictable, others only weakly so, still others not at all; that some facts are known quite directly and others only indirectly by chains of reasoning; and that the world, while showing a kind of order, is full of surprises. These basic generalizations about our epistemic condition can be reached by any thoughtful person and do not require *isolation*, *control*, and *manipulation* (= experiment, in a place called a "laboratory") or measuring instruments and explicit statistics. This sort of reliance on common experience when analyzing human knowledge was not only characteristic of those labeled "empiricists" by the historian of ideas (e.g., Locke, Berkeley, Hume, Mill) but is found also in the Continental "rationalists" (e.g., Descartes, Leibnitz, Kant). It is therefore misleading when some philosophers of science write as if "naturalized epistemology" and "first philosophy" are totally disjunct, implying that one who conceives metatheory as I do in this article must avoid the kinds of reasoning that regularly occur in Western epistemology from Descartes to Locke to Mill to the "moderns" Carnap, Feigl, Popper, Pap, Reichenbach, Salmon, and Co.

[4]I believe the only scholars who have clearly advocated (and begun to explicate) a strong form of this approach are David Faust and myself, both clinical psychologists (Faust, 1984; Faust & Meehl, 1992; Meehl, 1983, pp. 371-372 [1991, p. 303], 1990a, 1990b, 1990e, 1992b). This seems odd, but the sequel explains why. Probably Laudan (1990) should be counted as agreeing with an actuarial approach, as he says "empirical information about the relative frequencies with which various epistemic means are likely to promote sundry epistemic ends is a crucial desideratum for deciding on the correctness of epistemic rules" (p. 46). Similar remarks have been made by Diamond (1988, p. 181) and Feyerabend (1987, p. 171). Historian Frank Sulloway (1990) of M.I.T. is applying statistics to predicting individual scientists' early acceptance of correct theories, but he focuses on personal attributes of scientists (e.g., age, religion, birth order) rather than on the "objective" properties of successful theories. Psychologist Dean Keith Simonton, in studying scientific and and other kinds of creativity, emphasizes that subjective judgments of experts are often flatly contradicted by the most basic cliometric treatments (Simonton, 1990, pp. 89, 99-100, 144-146).

be more empirical in reliance upon the history of science than it was as they pursued it. Some of the younger generation portray the logical positivists (or, more broadly, the "empiricists") of 1920-1960 as more simplistic than they were. I knew many of them, some of them very well, and they were not stupid people! But it is certainly true that they relied on selected episodes from the history of scientific change "primarily as illustrative rather than probative" (Donovan, Laudan, & Laudan, 1988, p. 4). The story of Einstein and the 1919 eclipse (somewhat idealized, Earman & Glymour, 1980), Mendeleev's table, the kinetic theory of heat, the refutation of phlogiston and caloric, a bit of Darwin or Mendel were often recounted in lectures and writings by that older generation.

However, after the initial intoxication of "dismantling positivism," and the increased reliance upon history of science as the basis for developing a satisfactory philosophy of science, matters have become somewhat murky.[5] As Donovan, Laudan, and Laudan (1988) point out, in recent years "the historical school of theorists of scientific change … has shown signs of losing momentum, largely because no serious attempt has been made to determine the extent to which relevant evidence supports the claims made by the various theories that have been proposed—an ironic situation in light of the importance this school attaches to the empirical grounding of theories of scientific change" (p. 6). The neo-positivist "received view" has not, alas, been replaced by a high consensus paradigm; rather there is an increased divergence of positions, to the point that there is not even agreement as to the *aim* of metatheory, or as to what sort of discipline it is (social science? formal? statistical? language analysis? something else *sui generis*?). One need not be a pessimist or frustrated dogmatist to conclude that philosophy of science looks pretty much like a degenerating research program (Lakatos, 1970). Doubtless there are multiple causes for this, and I address only one of them. It is my belief that continued reliance upon the case study method as the sole way of assessing metatheory will lead to an interminable disagreement of the same kind that has existed for a century now among clinical practitioners and theorists of psychopathology, and for similar epistemological reasons.

PRESUPPOSITIONS

Before arguing for what I call the Strong Actuarial Thesis, espoused by me and by David Faust (see Faust & Meehl, 1992), it is necessary to state briefly my presuppositions about the metatheoretical enterprise, which I shall do with a minimum of argumentation because I expect that this much of my views is generally acceptable.

[5]An exception is those social scientists who, having an ax to grind against positivism, take everything Thomas Kuhn said as gospel truth. Speaking for my own field (clinical psychology), I regard Kuhn's (1970) book with ambivalence. I observe that those clinicians and personality theorists who chafe under the burden of proof for their assertions employ the name of Kuhn as a club against critics; or they use it to reduce their own guilt, to the extent they are aware that their theory and practice are not legitimated by evidence of the kind that statistical and experimental psychologists demand. I could say a lot about obscurantist motives in the social sciences for adulating Kuhn, but that's a topic for another paper.

I take it for granted that metatheory should try to get whatever help it can from other empirical and formal disciplines. These include cognitive psychology, meta-analysis (rapidly replacing the traditional narrative summary of research literature in the social sciences), and the clinical versus statistical issue, one of my fields of expertise as a psychologist (Meehl, 1954/1996, 1973, 1986; Dawes, Faust, & Meehl, 1989).

However, unlike some of the younger metatheoreticians, I do not anticipate that the empirical part of the cognitive sciences (perception, learning, problem solving) will be helpful except in their *negative* aspects, i.e., ascertaining which tasks the human mind does rather inaccurately and clumsily, and why. I conjecture the cognitive sciences that will be helpful in positive ways are those which are more formal, i.e., largely logical and mathematical, including artificial intelligence (AI) (Meehl, 1990a, 1990b).

I take metatheory to be the rational reconstruction of the history of science. Like any other theory, it will contain both factual and formal elements. Like other theories of human conduct involving intentionality, it will contain *descriptive* and *prescriptive* components, related in subtle and complex ways. For example, in the social sciences some think of "decision theory" as the empirical study of how organisms in fact select alternative courses of action. But even tough-minded, "nonphilosophical" empiricists, in studying animal behavior or human economic behavior, find themselves willy-nilly introducing prescriptive, normative elements. The white rat pressing a lever in the Skinner box adjusts its rate of responding to the reinforcement probability; and the different kinds of cumulative records that are produced in that experimental context are readily identifiable by an undergraduate who has been shown a few paradigm graphs. We do not suppose that the rat is an infallible decision maker, but it is true that on, say, a VR reinforcement schedule, the rat maintains a steady state of responding, putting out a linear cumulative record with a slope nicely adjusted to the reinforcement probability and also varying with the hunger drive. A behaviorist does not feel guilty of "mentalism" or "anthropomorphism" when he points out that this is a biologically sensible thing for the rat to do; but neither does he formulate these laws in a strong teleological form, which would require that the rat always "adapts" or "adjusts successfully" so as to stay alive (cf. Meehl, 1962a, 1992c). I shall say more about this and similar examples below, relating to the descriptive/prescriptive distinction.

The logical positivists had a phobia about the sin of psychologism, doubtless one reason why they did not pursue metatheory at an empirical level despite their commitment to empiricism. But one gets the impression that things have swung a bit far to the other side in the younger generation of metatheorists, that they are somewhat

phobic about invoking rationality or "logicalism" (I can't use ⌜logicism⌝, as I have been preempted by the metamathematicians). I don't understand why a metatheorist who views metatheory as the rational reconstruction of the history of science should have such reluctance. Other empirical, nonmetaphysical, nonphilosophical disciplines—so far as I know, all of them, from botany to astronomy—take it for granted that logic, mathematics, probability theory, set theory, and the like are legitimate analytical tools, part of the conceptual machinery that one relies upon in studying the empirical world.

When we study the diagnostic behavior of psychologists or physicians in a domain where they perform inefficiently as objectively evaluated by the patient's subsequent course or by the findings of the pathologist (Meehl, 1973, pp. 225-302, Chapter "Why I do not attend case conferences"), we permit ourselves to invoke Bayes's Theorem in the course of analyzing the sources of their inefficiency. But Bayes's Theorem is a piece of formalism, not an empirical law about how the human mind works. [The mind's inefficiency is partly due to that discordance (Kahneman, Slovic, & Tversky, 1982; Dawes, 1988).] I don't wish to belabor the point, I only want to be clear that I shall presuppose that the rules of logic and mathematics, including a subset of those concepts that we usually think of as belonging also to "inductive logic" and "epistemology," are legitimate concepts in doing metatheory.

An adequate metatheory must explain why science does better than other disciplines that make cognitive claims, but it should also explain scientific failures and inefficiencies. Any metatheory purporting to show that science always succeeds and succeeds as fast as it possibly could, must be false. Any metatheory that does not show why science progresses, why it tends (in the long run) to command the assent of all rational informed persons, must have something wrong with it. If it should happen that a reader does not believe that Watson and Crick's theory of the gene or Mendeleev's table is in better shape than Jung's theory of neurosis or Spengler's theory of history, I simply do not speak to his condition. Whether we reflect on its coherence and elegance, its explanatory and predictive power, its ability to command the assent of almost all honest inquirers, or its technological efficacy, physics is a more impressive cognitive enterprise than psychotherapy, not to mention such "soft" disciplines as esthetics, ethics, theology, metaphysics, jurisprudence, politics, literary criticism, or psychohistory.[6]

Suppose the metatheorist succeeds in distilling out from the habits, attitudes, and practices of scientists what it is that they do (and, perhaps even more importantly,

[6]This rather obvious comparison of scientific status should not be taken as denigrating these "soft" disciplines. I have made part of my living at psychotherapy for half a century. But a psychotherapist who claims it is a *science* is simply deceiving himself. It is an art, and an art for the most part not even based upon a science.

what it is they don't do!). These generalizations, which *can* be viewed purely descriptively as the cognitive habits that tend to scientific success, also lead directly to prescriptions: "He who wants scientific success should do the kinds of things that tend to achieve it." But formulations of successful scientific practice cannot have the character of rules in the narrow sense of that word; rather they are "general policy" or "guidelines," they are like the jurisprudent's *principles* rather than *rules* [Dworkin (1967) adumbrated by Roscoe Pound (1959), who in turn attributes this dichotomy to the Austrian jurisprudent Von Jhering; see also the reply to Dworkin by Christie (1968)].

There seems to be some difficulty among metatheorists about this "guidelines" concept. For example, we have the clash between a Popperian who cites examples of quick falsification and a Feyerabendian who likes to dredge up the (dramatic because atypical) history of Prout's Hypothesis. I must confess I do not understand what the difficulty is because the proper paradigm or analogy here is not an algorithm for solving a quadratic or taking a derivative but rather the *policy* of buying life insurance if you have a spouse and six children or the policy of betting on the best horse, or first trying the treatment of choice in medicine. One does not reject open-heart surgery for a life-threatening coronary artery problem on the ground that *sometimes* people die on the operating table. One does not say in retrospect that it was stupid to bet on Bluenose to place in the third when Bluenose happens to break his leg coming around the last bend. A man who buys life insurance and lives to age 106, surviving his wife and six children, may be said, in a sense, to have "lost a bet with the insurance company"; but we do not say, even after the fact, that he did something irrational. There are two meanings of ⌜error⌝, and not all failure to get at the truth represents error in the narrow sense of having done something irrational, stupid, following a bad policy, or making a logical or mathematical mistake. As Mark Twain put it, the essence of a horse race is a difference of opinion.

The aim of the metatheorist is understanding how science works, but it is obvious that if one considers science to be a success, a person who opts for playing the game of science cannot rationally ignore whatever distillation of effective scientific practices the metatheorist succeeds in discerning. It is an interesting sidelight of the logical empiricist movement that most of them—all those I have read, and those that I knew personally and pressed about this—were skittish about the notion that they might say something prescriptive to the scientist. Even as an undergraduate, talking with Herbert Feigl (in 1940), this struck me as a rather odd position to take. I think it arose from the logical positivists' great admiration for science and their contempt for speculative metaphysics (a là Hegel and Co.), doubtless admirable attitudes. But inasmuch as they insisted that philosophy of science was distinguishable from pure history of science, I found the

position incoherent and I still do (Meehl, 1984). One may say that God did not make man two legged and leave it to Aristotle to make him rational; but to the extent that Aristotle succeeded (with his rules of the syllogism) in distinguishing valid inferences from formal fallacies, it would be strange to say modestly that we ought never to tell anybody who offers a fallacious argument that he has committed an Illicit Major or an Undistributed Middle. Whether the distillations of metatheorists from studying successful and unsuccessful scientific histories are misformulated as strict rules (automatic truth-grinding machines) or are properly formulated as guidelines (general principles, pieces of friendly advice to the scientist), I don't see how anyone who distinguishes between metatheory and ordinary (nonphilosophical) history of science can deny that there is an unavoidable prescriptive element, however tentatively and humbly we put it when advising the scientist. In my work as a psychologist, I have never felt offended by advice from philosophers and can easily name two dozen whose metatheoretical criticism was helpful to me.

The hope would be that the majority of practitioners of "normal science" will take the advice of the metatheorist, becoming persuaded that he makes a good case for a certain set of guidelines; that even the innovators who make scientific revolutions will tend, by and large, to take such advice; and that the social pressure of the scientific community—not to mention fund granting agencies—will assure that only a few mavericks depart from policy. We *want* a few mavericks to do so, but we want the social pressure to be strong enough so that these persons will be courageous, autonomous, genius mavericks rather than merely cranks. We trust that the pressure of the understood policies will deter ambitious young physicists from attempting to invent perpetual motion machines but will not be strong enough to deter Barbara McClintock (a maverick in biology) from going her own Nobel laureate path (Keller, 1983; see also Margoshes & Litt, 1965). For a stimulating novel approach to the "epistemic optimizing" problem in a social group, see Kitcher (1990).

An adequate rational reconstruction should also enable us to look insightfully at the bad luck cases and the good luck cases when either goes against the long run adequacy of a guideline. One may safely presume that one way novel discoveries can be made in metatheory is by looking closely at these counter-statistical instances and tallying frequencies of aberrant properties found by case study.

At the present time, pronounced disagreements among philosophers of science persist despite the strong tendency for the current generation to rely more upon the empirical history of science, taking episodes as probative rather than merely illustrative. What can case studies be expected to tell us about a guideline or principle for appraising competing theories or for deciding what theoretical and empirical work may

be expected either to further a research program or to classify it as degenerating (Lakatos, 1970)? I do not argue that they tell us nothing, but I contend that even when carefully conducted by using the best canons of historical method to find out "Wie est eigentlich gewesen ist" *casewise*, they are not able to answer the policy question.

When advocating the adoption of a prescription or proscription, or when inferring that certain other theoretical properties falling under another principle ought to countervail it, one is making an *inherently statistical assertion*. In animal psychology the defects of the "anecdotal method" have been well known since the turn of the century, although some excessively skeptical psychologists have overdone it, especially in the context of discovery [see Reichenbach (1938) on the contexts of discovery and justification]. In clinical psychology we do not ordinarily speak of clinical case studies as "anecdotal" because of the pejorative flavor of that term, but, as I have argued elsewhere (Meehl, 1954/1996, 1973, 1983, 1987; Dawes, Faust, & Meehl, 1989), the defects of current clinical practice of diagnosis, prognosis, and treatment are in considerable part attributable to reliance upon the case study method in the absence of experimental studies or, when those are not feasible or applicable for reasons of generalizability, the statistical analysis of clinical file data. (For a frightening example from 20th century medicine, in which thousands of infants became blind iatrogenically because physicians were relying upon their clinical impressions rather than statistics, see the item on retrolental fibroplasia in the 1953-1954 *Annual Report of the* [British] *Medical Research Council*.) I want to emphasize that I am not here faulting the historian for using conventional historiographic methods to ascertain the facts about a particular episode. One objection to the anecdotal method in social science is that one frequently does not know, and cannot find out, some of the relevant circumstances—the idiographic particulars—surrounding a particular episode. Thus, the anecdotal method used by nineteenth century authors (e.g., Romanes, 1883) to prove the extent to which animals "reason" is flawed by our lack of knowledge of the rewarding and punishing experiences of our favorite dog or cat when outside our purview, or by our having perhaps forgotten such instances. That is not the problem here. Assume the historian of science *correctly* reconstructs a particular episode in the history of science in which a scientist proceeded in accordance with a certain methodological principle or guideline and "succeeded" (shown by the subsequent history of his experiment or modification of the theory). We may accept everything that is said about the episode including the scientist's motivations, how explicit the methodological principle was and why it was followed this time and not other times, and the like. But all it can tell us as a single episode is that it was an instance in which adopting the principle

in question was associated with a favorable outcome. In many cases we cannot even be confident that the favorable outcome occurred *because* the scientist conformed to the principle.

If we looked upon the anecdotal method in a Popperian way, we would be asking what purported methodological principle of "scientific method" does a particular case study *refute*? Employing this admittedly tough meta-criterion with controversial papers and books by Popper (1959, 1962), Kuhn (1970), Feyerabend (1970), and others, we find that the case histories offered by the various protagonists do serve to refute certain strong generalizations about how scientists proceed; *but the generalizations that are thus clearly and cleanly refuted by the anecdotes are generalizations that nobody has seriously maintained*. Take Feyerabend's favorite example against Lakatos's (1970) idea of a degenerating program, the Prout Hypothesis (that the atoms of all elements are "built up" from hydrogen atoms). For more than a generation of chemists, say before the American Civil War and up until the turn of the century (when the concept of isotopes was introduced), it was impossible to reconcile Prout's Hypothesis with the observed atomic weight of chlorine being 35.5, and there were hardly any supporters of it for half a century. Given the concept of an isotope, we discover that terrestrial chlorine is a mixture of two atomic weights with the same nuclear charge Z, hence the same electron configuration, and hence the same valence characteristics. Prout's hypothesis is thereupon revived and is today the received conception of cosmologists as to how the various elements were formed. What does this example prove against Lakatos or Popper? Or, for that matter, what does this example refute that was held by the Vienna positivists? It clearly refutes the empirical statement, "No scientific theory, abandoned because it was confronted by a persistent and 'central' recalcitrant fact (anomaly), has ever been subsequently resurrected in the light of altered auxiliary theory or new facts." Has anyone ever asserted this strong universal negative thesis as an empirical thesis in the history of science?

Shifting from an empirical generalization as to what scientists historically have in fact done to the prescriptive mode, I doubt that any metatheorist, even the most rigidly positivistic, ever laid down as a methodological dictum, "If what has appeared to be a definitive falsifying fact turns out (in the light of new evidence, or an amended theory, or different auxiliaries) no longer to be a falsifier, and if a body of evidence otherwise has spoken strongly in favor of the apparently falsified theory, it is nevertheless forbidden to resurrect it." One hardly needs a literature search on the older philosophers of science to be confident that no such proscription would be found.

What about sticking to an otherwise "good theory" despite an apparent recalcitrant fact? Mendeleev obviously had a lot going for him in the periodic table, so that the

incorrect proposed placement—relying on the then accepted (but incorrect) atomic weights of gold and tellurium—of an undiscovered element did not discourage him. He said that the "facts" must be wrong, as they later turned out to be. What does this important episode in the history of chemistry tell us? It refutes the claim that "No scientist ever stuck to his guns in the presence of what appeared to be a clear falsifier, and turned out to be right, nevertheless." Has any historian of science or metatheorist ever held this universal generalization? I know of none. You might say that, while nobody has ever said anything this strong about the empirical history of science, Popper for one has said that there *should* not be any such episodes if scientists proceeded properly. I will concede that is the over-all thrust of Popper's message, but, as Lakatos (1970) points out in defending "tenacity of a theory," Popper said in 1940 (and again in 1957) that what he labels "the dogmatic attitude" of "sticking to a theory as long as possible" is sometimes necessary in order to "*find out* its strength" (see Popper, 1962, pp. 49, 312). That early date suggests that Popper did not need the criticisms of his 1935 book (largely neglected and not translated from the German until 1959) to include this important buffering of his falsificationist thesis. I believe Popper himself has used the phrase *theoretical tenacity* non-pejoratively but cannot locate a passage; perhaps it was in a personal communication (1962-1963) when he visited the Minnesota Center for Philosophy of Science. Whether this concession to scientific practice should be considered an "undigested anomaly" in his metatheoretical research program (as Lakatos suggests, 1970, p. 177, fn 3 [1978, p. 89, fn 5]), I will not discuss except to say that Lakatos's "Popper$_0$, the naive falsificationist," never existed, as Popper shows by quoting from his earliest writings (Popper, 1983, pp. xxi-xxv). Two short and easy answers would be: First, an honest scientist may have stated what *would* constitute a falsification of his theory if everything else in the purportedly falsifying equation were taken as unproblematic but choose to stick with the theory knowing that these other matters are *not* certain.[7] A second possibility is that the scientist may take T to have been falsified *as it stands*, literally, but conjecture that some modification $T \rightarrow T'$ of its peripheral ("noncore") postulates would be successful. In examining Popper's position, one must be careful to distinguish *falsification of T* from *abandonment of T*, the decision to "cease working on it" (cf. Watkins, 1984, pp. 156-159). Neither of these tactics commits a

[7]The *modus tollens* falsification has major premise

$$T \cdot A_t \cdot C_p \cdot A_I \cdot C_n \vdash (O_1 \supset O_2)$$

where T = theory being tested, A_t = auxiliary theories presupposed, C_p = ceteris paribus clause, A_I = instrument auxiliaries, C_n = particular conditions allegedly realized by experimenter, O_1 and O_2 = observations made (see discussion in Meehl, 1990a, 1990b). Falsification of T is avoidable by not admitting O_1 and/or $\sim O_2$ into the corpus or by challenging one or more of the four left-hand conjuncts A_t, C_p, A_I, C_n.

logical inconsistency or methodological "bad faith," when *T* has previously accumulated "money in the bank" by making successful risky predictions (Meehl, 1990a).

In his *Realism and the Aim of Science*, Popper, responding to critics of his falsificationism, says that although he has not had time to survey the history of physics systematically, he does not doubt that there are hundreds of examples of theories being falsified quickly and clearly, and he provides "a list of [20] examples chosen almost at random" (1983, p. xxvi-xxx). These examples corroborate the existential statement, "Cases have occurred of a theory being clearly and quickly falsified by experiment and promptly abandoned." What do these twenty examples *refute*, assuming each of them to be historically accurate as Popper states them? They refute the statement, "No scientific theory has ever been quickly abandoned because of what appeared to be a clearly falsifying fact." Does Lakatos deny this? I can't find any such denial in his writings. What Lakatos argues is that there are numerous examples where theories are *not* abandoned on this basis and that it is always logically possible for a theorist to preserve the theory in the presence of such an apparent falsifier—a seeming truth of formal logic relied on by Quine and Duhem but challenged by Grünbaum (1960).

It is needless to pile up examples. The basic point, once stated, seems obvious even without citing the literature: Instances in which strong methodological "rules"—whether positivist, Popperian, or whatever—have been successfully followed occur in profusion as do instances in which such rules have been successfully ignored. Given the ampliative character of empirical science (induction, if you like, or bold conjectures with an attempt to falsify, if you prefer), this is precisely what one should expect. The enterprise of empirical science involves working in ignorance of the way things really are, relying upon *samples* of the way things act when we manipulate (or selectively attend) in certain ways. The theoretical scientist is sampling from a huge domain of actual and possible data relations (Meehl 1990b, pp. 13-14, 39-42). Like someone buying life insurance, or betting on a horse race, or deciding whether to propose marriage or where to take graduate work for a PhD, the scientist is drawing conclusions with incomplete information about the state of nature. Putting it simply, not all possible experiments have been performed, and not all real events have been observed and recorded. Like the psychotherapist interpreting a dream, or the internist diagnosing a difficult case, the research scientist may act with the greatest rationality that the knowledge situation permits and yet fail. A scientist may also act in a way that most informed persons would consider irrational and nevertheless succeed. Why should this puzzle us? In any domain in which the relationships among a finite sample of all possible facts that are causal consequences of the state of nature are intrinsically stochastic, the inevitable state of

affairs is a four-fold table. One dichotomy is that the would-be knower can either proceed in accordance with "best long run policy" or refuse to do so; and the other dichotomy is that of cognitive success or failure. In an intrinsically stochastic situation, which nobody denies is the character of inductive inference or the methodology of the empirical sciences, none of the four cells of the table is empty. *It is obvious that in all human doxastic enterprises, whether formulated as dichotomous decisions to accept or to act on "as if accepted," or as assignments of degrees of rational belief, or as rough appraisals of verisimilitude—it doesn't matter how you put it—no procedure will guarantee success, and no departure from a statistically optimal procedure will, in a particular case, guarantee an instance of failure*. Is there any statistician, logician, or historian of science who would dispute this statement? I think not.

Given these understandings, it seems obvious that metatheory can only aim to formulate policies, general principles, guidelines, rules of thumb, suggestions, "helpful advice" with the explicit understanding that such advice will not always result in cognitive attainment, and with the hope that a subset (not too large in number) of theorists (ideally brilliant ones) will depart from the advice. The most that can be claimed about the best such advice is that it rests on the *statistical* claim that proceeding that way *tends* to pay. As with principles of jurisprudence, there are various principles that can under certain circumstances countervail each other (Mr. Justice Black said the Supreme Court had to perform a "balancing act" in such cases); but since they are principles rather than rules, they cannot, strictly speaking, *contradict* each other.[8] A meta-principle about such countervailings (comparable to a decision rule between two *prima facie* moral obligations in the field of ethics, or two accepted principles of economic or political policy) will itself have to take account of actuarial success rates.

THE HUMAN BRAIN IS IMPERFECTLY RATIONAL, A FALLIBLE TOOL BY NATURE AND BY TRAINING

If we grant that a set of guidelines for conducting scientific theorizing and research successfully should be distillable from historical studies of episodes in the history of scientific change, how should the metatheorist proceed in investigating this? The received view, which seems universally held, is that one does it by the case method. Here is where I think the cognitive sciences, with an assist from the fields of social and clinical psychology, have something helpful to say. There is a fairly large and consistent body of evidence which shows that the human mind is not as good as we might

[8]In their fascinating Talmudic collection of several metatheorists' *descriptive* generalizations about scientific practice, the Virginia Polytechnic Institute group (Laudan, Donovan, Laudan, Barker, Brown, Leplin, Thagard, & Wykstra, 1986) show that there are numerous "contradictions"—even in the same metatheorist's writings—if we treat the principles in a yes-or-no all-or-none manner, i.e., as rules. Their paper provides one of the strongest arguments I know for the actuarial approach.

have hoped in evaluating even small and simple sets of data, let alone large masses of complicated information, and in quite a few situations is discouragingly poor at it. I shall not attempt to summarize that research here but merely refer the interested reader to literature on clinical versus statistical prediction in psychopathology, educational selection, and other situations (Dawes, Faust, & Meehl, 1989; Meehl, 1954/1996; Sawyer, 1966; Sines, 1970; Wiggins, 1981), and to studies showing that even in simple inference situations with a small number of variables the human brain is not a good strategy selector, an effective data summarizer, a rational evidence assessor, or an accurate assigner of statistical weights (e.g., Arkes & Hammond, 1986; Dawes, 1988; Faust, 1984; Hogarth, 1987; Kahneman, Slovic, & Tversky, 1982; Kleinmuntz, 1990; Nisbett & Ross, 1980). Table 1 lists some sources of error in clinicians' diagnostic and predictive judgments (completeness is not claimed) with a parallel column showing similar or identical factors working against optimal rational judgments by scientists about theories.[9]

Ever since writing the foreword to Faust's (1984) book, I have been especially alert to symptoms of irrationality in the discourse of scientists. I can present no statistics, but I do offer a conjecture, as confidently as one can or should from non-tallied, impressionistic "data": I predict that a formal statistical content analysis of scientific communications (books, articles, letters to editors, conference talks, media releases) dealing with controversial matters would reveal a rather high incidence of poor reasoning, ranging from mere "weakness" of case and slanted semantics to misstatement of facts and grossly fallacious arguments. Such an actuarial study would be easy to do, employing as judges uncommitted scientists familiar with the topic and logicians who have been given basic instruction in the scientific domain. Select domains whose core concepts and mathematical formalism are not too forbidding to bright persons of generally "scientific" education. Judges should receive rating instructions, followed by corrective feed-back after initial rating trials (data discarded), to reduce the usual rating errors [e.g., halo, central tendency, leniency, bias, strain toward consistency (Guilford, 1954)]. "Personal equation" calibration and pooling should provide adequate dispersion and Spearman-Brown boosted reliabilities. (I remind the reader that pairwise interjudge reliabilities of only $r = .60$ suffice to achieve a pooled 7-rater reliability $r = .91$.) If factor analysis showed some judges persistently superior to others, ratings could be weighted

[9]Concern that judgment experiments in the laboratory may have low ecological validity, because of evolutionary adaptive considerations as to "natural environments," has been rebutted by Arkes (1991). That objection, even if valid, would not apply to the large mass of studies on clinicians' decision making in "real life" practice or to the similar findings on business executives, physicians, parole officers, sports writers, academic selection committees, military personnel officers, and so on.

TABLE 1
SOURCES OF ERROR COMMON TO CLINICIAN AND SCIENTIST

Clinician: Error Sources	Scientist: Error Sources
I. Objective: Would operate on ideal clinician or ideal scientist	
Facts known about patient and situation are only a sample: (a) Biased and (b) random error	Experiments performed are only a sample of possible arrangements-*cum*-observations: (a) Biased and (b) random error
"Facts" not equally trustworthy Source (informant, documents): Biased? Track record for accuracy known? How many sources agree?	Experimental results not equally trustworthy Experimenter: Biased? Track record for replicability known? How many replications?
No inductive logic algorithm exists (except Bayes's Theorem in special cases)	As with clinician
Facts are not independent, hence their mutual corroboration is weakened in various degrees	As with clinician
Two-way relationship (T ↔ F) between theory and facts Which direction should control in a given instance of incompatibility?	As with clinician
Time pressure to decide (e.g., interpret, reflect, or remain silent?)	Rarely a factor, except for priority motivation or grant application deadline
Causality/correlation problem when manipulation not feasible, must "take data as they come"	As with clinician
Facts not obtainable (too costly, too risky, in the past, not reported)	
Cannot safely assume absence of a rare factor, "other things being equal"	As with clinician
Cannot safely assume statistically "normal, usual" conditions prevail	*Ceteris paribus* clause problematic
Psychological tests often have low, moderate, or unknown validity	Auxiliaries problematic
Episodes subsumable under different traits (e.g., large tip to waiter: generosity? showing off? impulsiveness? drunk?)	Some scientific instruments have low, or unknown, precision Experimental result plausibly explainable by different theories
II. Subjective: Correlated with personal traits of clinician or scientist, pressures to deviate from rational ideal, and some general to all clinicians or scientists	
A. "Neutral" (unmotivated) Cognitive Error	
Suboptimal or even nonsatisficing inferential "strategies," common bad habits, biased "heuristics" (e.g., availability, recency, vividness, neglected baserates) as described by Arkes and Hammond (1986), Dawes (1988), Faust (1984), Hogarth (1987), Kahneman, Slovic, and Tversky (1982), Kleinmuntz (1990), Meehl (1973), Nisbett and Ross (1980)	As with clinician
Premature freezing of diagnosis or interpretation	Premature commitment to a theory
Selective recall of episodes	Selective recall of experiments; inadequate literature search
Inertia, preference for the familiar, avoidance of having to revise ideas or think hard	As with clinician

(continued on next page)

TABLE 1 (CONT'D)
SOURCES OF ERROR COMMON TO CLINICIAN AND SCIENTIST

Clinician: Error Sources	Scientist: Error Sources
Seeking confirmation rather than refutation	As with clinician
Failure to scan all possibilities	As with clinician
Addition of fact f_2 as makeweight. Already have relied on fact f_1, and general theory or observational linkages and constraints entail $f_1 \rightarrow f_2$ with high probability, thus f_2 adds little or no new support	As with clinician
Over-reaction to recent bad diagnostic error (e.g., suicide), or to a clever correct one	Over-reaction to falsification of a favorite theory, or an ingenious experimental corroboration
Imperfect (informal, subjective) computation of utility × probability	As with clinician
Assignment of nonoptimal weights to facts	As with clinician
Inconsistent application of weights ("unreliability," Goldberg paradox)	As with clinician
Formal reasoning error (logic, mathematics)	As with clinician
Computational error (e.g., biggest source of unreliability for the Stanford-Binet is in scoring)	As with clinician
Information overload: Human mind cannot store and process large mass of information about a patient	As with clinician
Hindsight bias: mistakenly claiming that one would have predicted a (now known) disease, symptom, or life-history fact if one had not been informed of it (Arkes, Wortmann, Saville, & Harkness, 1981; Dawes, 1988; Hawkins & Hastie, 1990)	As with clinician (Slovic & Fischoff, 1977)
B. Motivated Error	
1. Idiosyncratic, personal, psychodynamic	
Projection (naive or defensive) of traits, themes, motives, deficiencies	n.a., except for naive projection, "surely everyone agrees that …"
Denial of traits, themes, defenses, etc.	n.a., except for aesthetic or cognitive distaste for certain kinds of concepts, experiments, instruments
Identification with one's teacher, advisor, analyst, supervisor, school	As with clinician
Rejection of one's teacher, advisor, analyst, supervisor, school	As with clinician
Transference problems generally	As with clinician
Aggression (and reaction-formations against the impulse) Shock? Interpret? Tough love?	Aggression, dominance, intolerance of disagreement
Grandiosity, guru omniscience fantasy	Publication compulsion, ambition, spiteful competitiveness, envy of famous scientists
2. Socially acquired, shared, and maintained	
Ideology (political, religious, economic) (e.g., Lord, Ross, & Lepper, 1979)	As with clinician
Class interest (or gender, age, race, ethnic, nation, profession)	As with clinician
Theoretical (e.g., Freud, Skinner, Jung, Adler, Ellis)	As with clinician

proportionately to judges' first factor loadings in the pooled judgment. The scientist judges should have some instruction in logic and philosophy, including such (relatively) noncontroversial principles as the Total Evidence Rule, the concept of different interpretations ("models") of a formal calculus, the two meanings of probability, the valid and invalid figures of the implicative syllogism, the bidirectional "control" of theories by observational protocols (collectively, long run) and candidate protocols by theories (singly, short run), the crucial role of (often problematic) auxiliary theories in theory testing, etc. It *cannot* be safely assumed that being a "successful working scientist" guarantees this kind of basic logical competence in our raters, an assumption that would beg the very question we want to investigate. Research on transfer of training should prevent psychologists from such a mistake—but those studies themselves suggest that even we may not be so immunized when we shift from the roller skating/ice skating paradigm of transfer effects to the scientist observing/scientist theorizing situation.[10] In the training of scientists there occurs *repeated*, *explicit*, and *specific* indoctrination about observation and calculation. Undergraduates are warned about such things as parallax in reading an instrument dial, the concave and convex meniscus in a graduated tube, "centering" the white rat in maze entry, masking white noise, and clean test tubes. Statistical computations and mathematical derivations must be checked in various ways. By a combination of *formulated general principles* and *repeated concrete exemplifications*, the scientific mentor tries to reduce observational, mathematical, and numerical mistakes. But no comparable attention is (typically) paid to the difficult art of interpretation. It seems generally assumed that "how to observe accurately" and "how to compute correctly" are cognitive skills that must be taught, whereas "how to think clearly," even when the inferential structure is complex, is a skill the student will somehow pick up easily, informally, along the way. There is no good reason for assuming this, and what we know about transfer of training goes against it. In psychology, it may be that departments vary widely in this respect. Clinical psychology students on internships

[10] In addition to the research on resistance to scientific discovery (Barber, 1961), and nonrational "social" influences in the conduct of research (Brush, 1974; Goodstein & Brazis, 1970; Greenwald, 1975; Greenwald & Pratkanis, 1988; Greenwald, Pratkanis, Leippe, & Baumgardner, 1986; Latour & Woolger, 1979; Mahoney, 1976, 1979; Mitroff, 1974; Mynatt, Doherty, & Tweney, 1977, 1978; Rousseau, 1992; Tweney, Doherty, & Mynatt, 1981), we have such amazing findings as Mahoney and Kimper's (1976) that 39% of physicists and 67% of biologists cannot even recognize the *modus tollens* [$p \supset q$, $\sim q$, $\therefore \sim p$] as a valid syllogism! In a replication, Kern, Mirels, and Hinshaw (1983) infer that formal logic does not matter much in science, as shown by their research, demonstrating competent scientists not recognizing the falsifying figure (*modus tollens*) as valid. We see here the grip of the common view that, since science "succeeds," scientists must be doing fine just as they are! Of course, we have no good estimate of how many bad tactics, poor strategies, and erroneous decisions these admittedly successful scientists are committing by being insufficiently sensitive to the logical structure underlying Popper's theory of science.

have reported that other interns quickly learn that if they have a "methodological stomach ache," members of the peer group from Psychology Department X are the ones to consult.

Here is a simple example: In watching for nonoptimal thinking I have noticed a nearly universal looseness in the semantics of the crucial metaterm ⌜proof⌝, surely one of the most pervasive and important words a scientist uses in discussing theories. Scientist A presents a certain experiment or theoretical argument favoring (or opposing) theory *T*, and Scientist B comments "Well, that's interesting, but it's not proof of *T* [or, against *T*, as the case may be]." I find that the phrase ⌜proof of⌝ is used in at least seven ways. A fact F is said to be ⌜proof of⌝ (or ⌜proof for⌝, or ⌜tends to prove⌝, or ⌜is probative of⌝) a theory *T* if:

F is consistent with *T* (i.e., F does not *refute T* as an empirical finding ~F would have done);
F is relevant evidence for *T* (i.e., F supports or confirms *T* to some extent, the probability of *T* given F is distinguishably greater than it would be without F, or given ~F;
F renders *T* sufficiently probable to warrant (discretionary) credence, or enough to "work on *T*" further;
F makes *T* more probable than not (as in a civil litigation, $p \geq 50$);
F supports *T* so strongly that, absent counter-considerations, it would be irrational not to adopt it;
F supports *T* "beyond a reasonable doubt" (the criterion in criminal trials);
F makes *T* certain.

These seven meanings treat *T* alone, but further meanings arise in the (usual) situation of comparing theories (e.g., before we knew F, T_1 was ahead of T_2 on the evidence, but with F conjoined, T_2 has moved ahead).

I find that scientists *almost never* make clear which meaning of ⌜proof⌝ is intended, and the differences are not minor. A reader who is careless or naive may be badly fooled, or if himself biased, abuse the semantics so as to take what he wants from such a passage. The danger here about *factual* statements' probative value obviously exists also for theoretical, mathematical, and methodological *arguments*. I dare say many promising developments (especially by unknown or heterodox scientists) have been crushed by a prestigeful critic's polemical use of the term ⌜proof⌝. We may be sure there are other potent, loaded metatheoretical terms that suffer similar ambiguity or vagueness and are employed carelessly or tendentiously to the detriment of scientific thinking.

The easiest way to convince oneself that poor argumentation often occurs among scientists is to read about current theoretical or factual controversies in domains where one has no personal predilection. I, for example, have to watch for my biases as a clinician or former learning theorist. But I have never had much interest in the field of perception, so I can peruse a controversial article on the geon theory of vision (Biederman, 1987) with fair confidence that I will not see "errors" that are not there. The same is true for my reading about cancer, cosmology, linguistics, nutrition, archeology,

and other sciences where I have no opinions. Semipopular treatments of current issues written for scholarly nonspecialists, such as appear in *Scientific American*, *American Scientist*, and *Science News* are a good source. We do not know, absent quantitative content analysis of discourse samples, whether defective argumentation occurs less often in the more exact, mature, powerful sciences (e.g., physics, astronomy) than in the life sciences, although this seems widely assumed. I can only record my anecdotal impression that the difference is slight, if it exists. Controversial argument in astrophysics, if one judges by quotations in semipopular books (e.g., Overbye, 1991), is sometimes as sloppy, biased, and personalized as anything one comes across in disputes about history, jurisprudence, ethics, or psychoanalysis. I suspect that the cognitive superiority of chemistry over psychodynamics comes more from the nature of the subject matter, the precision of control and observational instruments, and the deductive fertility of the formalism than from any great superiority as to "clarity" or "fairness" in chemists' thinking or attitudes. That, again, is appropriate subject matter for research in the Faust-Meehl program. There is some evidence that taking courses in psychology improves students' "rational thinking" more than taking one in chemistry (Lehman, Lempert, & Nisbett, 1988; Nisbett, Fong, Lehman, & Cheng, 1987). This should not surprise anyone who has taken both. The less developed state of social science tends to direct attention to methodological problems (conceptual unclarities, factual disagreements, disputed principles, epistemic aims, tempting fallacies) even in an elementary class; whereas in beginning inorganic chemistry, with its clear, unproblematic received paradigm, there is no occasion to spend time on metatheory. I do not believe there was more than five minutes of such ("the scientific method" simplified) in my college chemistry class.

I do have an impression of possible difference on the *reader* side. So much of social science writing is trivial and "nondemanding" that it is easy to become a superficial reader; thus when a psychologist in one of the "soft" fields has to read something a bit more complicated or aiming at conceptual precision, he does it badly.[11]

---

[11]I permit myself three examples from my own writing, because they are so clear—I would say *glaring*—that I do not worry about my bias. In 1948, MacCorquodale and I published a paper on hypothetical constructs and intervening variables. Our aim was to *distinguish between* them, not to advocate either over the other. Hypothetical constructs are more interesting to some psychologists than intervening variables, but the formers' explanatory surplus meaning, while a source of attraction, has the inferential "riskiness" attendant upon saying much more than generalization of the facts. We could not have co-authored a preference, because we differed strongly in that respect. Yet well over half of books, articles, letters, or conversations alleged that we were *for* (or *against*, depending on the bias of the reader!) hypothetical constructs. In my monograph on prediction (Meehl, 1954) I distinguished with meticulous explicitness (pp. 15-18) between *kind of data* (i.e., psychometric, life history, clinical interview, ward behavior) and *mode of combining data* (i.e., statistical formula or actuarial table versus clinical judgment) and listed examples of the various combinations that arise. I found that most readers completely ignored this crucial distinction and proceeded to conflate the dichotomies. Most published discussions of the book, from 1954 to the present, state "Meehl argues that mental tests are superior to the interview," an assertion found nowhere in the book and on which I in fact had—and still have—no settled opinion! In my theory of schizophrenia (Meehl, 1962b, 1989b, 1990c, 1990d) I conjectured that all persons inheriting the integrative neural defect (schizotaxia) "become, *on all actually existing social learning regimes*, schizotypes in personality organization" (Meehl, 1962b, p. 831 [italics in original]). Some critics complained that I should not have used two terms for denoting the same class of persons. My italicized phrase was simply ignored, not to mention the difference between a disposition and its activation, or the elementary distinction between meaning and reference. "All animals with a heart have a kidney" does not entail the synonymy "heart = kidney." From Charles' Law that, for

Suboptimal reasoning among scientists, however caused, renders the non-actuarialized case study weak as a test of metatheoretical principles. Stated qualitatively, as philosophers usually do, any such principle cannot be "valid" except *ceteris paribus* (i.e., when other such qualitative principles do not threaten to countervail it). Suppose with this understanding of their stochastic character, we restate all the qualitative principles as "factors weighing pro and con." Then we need to ascertain the weights and the mathematical form of their composite. Assume we had the ideal composite function, probably nonlinear and interactive (cross-product terms *at least*), as provided by The Grand Epistemologist—who is less than Omniscient Jones as to physical nomologicals and particulars but does possess the optimizing knowledge-function. Would the historian of science, conducting case studies, come across *deviations* from this function? Of course he would, not only by individual scientists, but by the whole scientific community, due to irrationality.

A closer analogy to the metatheorist's cognitive problem is found in the new methodology called meta-analysis (Glass, 1976; Glass, McGaw, & Smith, 1981; Hunter, Schmidt, & Jackson, 1982). The conventional way of appraising psychological theories by a "narrative" review of the research literature frequently leaves the reader almost as much in doubt after reading the review as before (Meehl, 1990e). The inventors of meta-analysis point out that a clinician or social psychologist cannot store, retrieve, and apply appropriate computations to a set of scores on tests or other behavior data on 300 subjects "in his head, informally," and consequently must have recourse to statistical methods to discern the relationships that obtain in the data set. Similarly, they argue, 250 research experiments on a particular theory or on competing theories, in which the studies differ as to populations sampled, exact character of the instructions, behavior output observed, methods of measurement, etc., cannot be efficiently processed by the human brain. For example, in evaluating efficacy of psychotherapy, one can easily concoct a list of two dozen factors that therapists of various persuasions agree might plausibly influence the results, such as age, sex, IQ, diagnosis of patients, years of experience of practitioner,

---

constant pressure on a gas, $V = KT$, one does not conclude that volume and temperature are the same concept!

profession of practitioner (M.D., PhD, MSW), number and density of interviews, therapeutic mode, follow-up time, and method of assessing change. I can personally attest, as a practicing psychotherapist who is also a research psychologist, that reading a conventional narrative review of studies of psychotherapy or even writing one myself (Meehl, 1955) usually leaves me in doubt what to conclude. It was possible for skeptics like Eysenck to claim for 30 years that there was no convincing scientific evidence that psychotherapy had any efficacy, and it is only since the introduction of meta-analysis that it is no longer possible to maintain this skepticism (Smith, Glass, & Miller, 1980). Meta-analysis takes the individual study as the statistical unit, and the mean and standard error of an "effect size" (ES) are computed. The ESs are analyzed over the aggregate of studies with respect to each of the various factors that might influence the extent to which psychotherapy is efficacious in producing change as well as the interactions of these factors in the Fisherian sense. Thus, for example, one can ask whether the effect size of psychotherapy provided concurrently with psychotropic medication is close to what would be predicted by summing the psychotherapeutic and drug effects or whether the presence of one of these interventions potentiates the effect of the other.

It is sometimes objected that meta-analysis is a crude, shotgun method that cannot substitute for the scholar's consideration of the special factors discernible by intensive scrutiny of the individual studies. This is a misinterpretation of the method; any factor that a reviewing scientist can reliably discern in some studies and not others is itself included in the list of factors that the meta-analysis quantifies.

I do not advocate meta-analysis as a way of evaluating scientific theory (Meehl, 1990a, 1990e), since (as its authors point out) it was devised as a method of evaluating the impact of interventions, as an *appraiser of technology* (e.g., modes of educational instruction, effect of school class size, benefits of various drugs, results of psychotherapy, influence of television on children's violence) and was not designed to evaluate substantive causal theories. Its defects for the latter are obvious, notably that (1) a theory is not taken to be better confirmed the bigger the ES, but rather on the basis of how close an ES is to a predicted value; (2) no predictive *risk* is represented by ES or its standard error; and (3) there are other properties of theories than empirical point or interval prediction that are relevant in their appraisal (see below, p. 387). I offer meta-analysis here only as an example of the kind of cognitive difficulty that the human mind encounters in concluding with confidence from the conventional narrative summary of empirical results. My point in the present context is that the accumulation of dozens or hundreds of case studies of scientific change will present the metatheorist with the same kind of cognitive difficulty that a psychoclinician experiences in trying to make sense of

the conventional narrative summary of research on a problem such as my theory of schizophrenia (Meehl, 1962b, 1989b, 1990c, 1990d), or Freud's theory of dreams (1900/1953), or Festinger's theory of cognitive dissonance (1957). I therefore propose that, given the factually stochastic character of meta-theoretical generalizations about scientific success, and hence the intrinsically statistical character of guidelines for theory appraisal as prescriptions suggested by the scientific history, such generalizations and guidelines should be investigated by explicitly actuarial methods rather than by the conventional narrative account of scientific case studies.

CASE STUDIES AND STATISTICS

The problem of what can and cannot be properly inferred from the case study method is a longstanding issue in the social sciences which I surely cannot hope to "settle" in the present article. Fortunately one does not have to adjudicate all of the aspects of that controversy to argue that case studies in the history of science should be summarized statistically for certain purposes, and the performance of theories should be quantified in a way that does not hinge upon one's causal (psychosocial) interpretation of single case studies. Neither of those contentions involves a rejection of case studies as a method of investigation, a position that I, as a clinical practitioner, would hardly take. There is one function of the case study in psychology and sociology that is not problematic so we need not discuss it here, namely, that the case study can be a rich source of hypotheses. Even the most toughminded statistical critics of the case study as a method of validating conjectures have readily conceded its role in the context of discovery. Those strongly actuarial social scientists who accept this but reject the case study *totally* in the context of justification do have a paradox to deal with; it could hardly be that case study is sometimes a fruitful source of theoretical ideas, yet is always incapable of yielding valid conclusions. We should qualify the usual statement by saying, "The case study can be a fruitful source of conjectures, although it is *not by itself usually sufficient* to function as a strong corroborator of them." To suggest that it has no evidential value at all leaves one without an explanation of how it can possibly be a fruitful idea generator.

Even this concession, I think, says too little in its behalf. Given generalizations in a science, if the facts of a particular case are clear (the almost purely "observational" statements found in the study being admitted into the corpus), a case study is obviously capable of falsifying certain generalizations, as pointed out above. As Bertrand Russell somewhere says, "A single occurrence of an event establishes its possibility." *Example*: A mini-theory in psychopathology that relied upon Clark Hull's theory of learning led to a theory-based attempt to define ⌜anxiety⌝ as "the conditionable component of the unconditioned response to pain." The impulse to this definition sprang from the Hullian emphasis upon primary drives in the same way as many Freudian definitions spring

from Freud's notion of anaclitic cathexis. A relatively small number of case studies suffices to falsify this conjecture, despite its appeal to reductionist psychologists and the fact that it is compatible with some data from the animal laboratory. (This was partly because until more recently, under the influence of clinicians, animal researchers relied heavily upon electric shock or other "physically" painful stimuli in studying escape and avoidance learning.) There is a strange neurological disorder, now believed to be hereditary in origin, found in children in which there appears to be a total absence of conduction of stimulation from the pain receptors (free nerve endings) to the perceiving systems of the brain. These children come to the pediatrician's or neurologist's attention because they suffer grave injuries when they do not respond normally to painful stimuli (e.g., a severe burn from failing to withdraw a hand from a hot surface). They can have severe skin lesions or broken bones without complaint. Careful clinical study of a few such cases makes it clear that these children, for whom pain as a stimulus modality is literally nonexistent, are quite capable of experiencing normal anxiety (e.g., fear of strangers, object loss from departure of mother, apprehension of social rejection, "normal social fear" such as stage fright, and the like). It is not necessary to do any statistical study or to quantify the intensity of the anxiety reaction because any clinician or layperson will readily agree that these children are perfectly capable of fear in various situations, which makes the originally proposed theoretical definition of anxiety unsatisfactory (see, e.g., Cofer & Appley, 1964, pp. 261, 586, 701; McMurray, 1955; Sternbach, 1963).

Although I can present no quantitative evidence, I believe there are reasons for expecting the deliverances of the case study method in the history of science to be considerably more trustworthy—both as sources for subsequent statistical treatment and as providing a basis for plausible causal inferences casewise—than is true in political history or in the study of individuals' psychopathology as carried out by clinicians. This being itself a matter for empirical research, I do not attempt to prove it here but simply list the plausible armchair considerations. First, the documents available in examining a scientific episode are regularly contemporaneous with the events (thoughts, conversations, meetings, and observations) because of the scientist's habit of immediate recording. Where the document involved is an experimental protocol, it is part of scientific ethics to write it down at the time. In a psychiatric case study there usually are no documents, and the clinician has to rely almost entirely on the patient's or informant's recollections of events long past.

Second, scientific habits of mind, attitudes about evidence and truth, and the high value placed upon fairmindedness and objectivity should lead the scientist's documents—even including correspondence with critics or with editors who rejected a paper—to be, at

least on the average, somewhat more accurate than the productions of mental patients or their biased and ambivalent relatives, or the verbal productions of politicians, preachers, publicists, generals, and the like. I realize that among the younger generation of metatheorists there is a tendency to poke fun at the idealized scientist, attributing to the logical positivists the insane idea—which, so far as I know, not a single one of them held—that scientists have no prejudices, loyalties, national identifications, spite, conceit, selfish interests, paranoid tendencies, or other "irrational" motives. This is a straw man and I shall say no more about it. (I think it reflects the current cultural cycle, which is Dionysian rather than Apollonian, and a certain "failure of nerve," especially among people hypnotized by Kuhn or unduly influenced by second-rate social science propaganda.) For my part, I do not understand how anyone who has read polemical papers by scientists or listened to impassioned argument in a panel or symposium at a scientific meeting could say that even these emotionally charged instances are of no greater objectivity, emphasis upon rationality, or avoidance of grossly fallacious appeals than what one typically hears from politicians on the floor of Congress or in a sermon by a television evangelist, or reads in letters to the editor from literate laypersons. It goes without saying there are some scientists who on some topics can only put on a pretense of objectivity. But the difference is one of degree, and I remain convinced that the degree difference is sizable.[12] In my own field, I have been struck with the fact that psychologists working in domains that have the earmarks of a genuine science (convergence of informed persons, episodes of clear falsification, cumulative replicable empirical results) are not the ones who invoke the Kuhnian theses. The Kuhn enthusiasts are almost uniformly—I can think of only two exceptions among my colleagues—in the "soft" areas, such as personality theory, social psychology, and the less scientific parts of clinical psychology. These are social scientists who chafe under the burden of proof and who feel greatly relieved of this unpleasant burden by being able to say, "Oh, don't ask me to prove these things; Thomas Kuhn assures us that all observations are theory infected, all you need to do to understand and accept my theoretical position is to undergo a Gestalt switch." Their quotations from Kuhn are highly selective. They like to talk about theory ladenness and that there are no objective facts but merely different ways

[12]Scientific objectivity, detachment, and fair-mindedness in rational discussion is a regulative ideal which *of course* none of us reaches, anymore than we attain flawless moral conduct, errorless decisions, or perfect health. But, as an ideal, it has profoundly influenced the cognitive and communicative style of educated persons. Even ordinary civility in disputation has been improved by our internalizing the values of post-Galilean science. Contemporaries with "failure of nerve" and cynicism about science should have a look at the polemical style of medieval and Rennaissance controversy, e.g., the obscene vituperation indulged in by such an honest, conscientious, high-minded, and cultured man as Sir Thomas More in his tract against Luther (Kenny, 1983, p. 51). Luther regularly wrote like that in controversies.

of perceiving the world; but they never quote Kuhn's opinion (1970, pp. viii, 15, 37, 160, 178-179) that it is doubtful whether the social sciences have developed even to the point of having a paradigm!

A third reason for expecting case studies in the history of science to be more trustworthy and illuminating, especially with regard to casewise causal inference, is the smaller number of relevant variables involved in most episodes. I may know that a certain scientist has a "father problem" with respect to his PhD advisor and consequently a prejudice against papa's theory, that he tends to identify with French rather than German science, and that he is fond of theories employing an esoteric formalism that he knows about and most people do not. Of course there are such cases of scientist bias and irrationality. But compare variables of this kind, impinging upon his otherwise rational cognitive functions, with the vast number of variables that historians have seriously considered in analyzing the outbreak of the Great War. They range from the cyclothymia and grandiosity of Wilhelm II, to the anxiety neurosis of Russian foreign minister Isvolski, to the decay of the Austro-Hungarian dual monarchy, to the longstanding hatred between Croats and Serbs, to the malignant influence of Rasputin via the neurotic Tsarina on (passive and possibly schizoid?) Tsar Nicholas II, not to mention the incompetence of the Austrian prime minister Berchtoldt, characterized by a former schoolmate as "having neither the intellect nor integrity to manage a factory of twelve hands." One historian may emphasize the threat posed by the expansion of the German navy in relation to the British, but a Marxist might view that as merely symptomatic of the growing competition of the German capitalist class with the world trade dominant industrialists of England. Aside from subjectivity and defensiveness by the participants, access to some of the variables is impaired by the absence of crucial documents. One of the first things that happens just before a formal declaration of war takes place (and the ambassador is instructed to turn over his passports and leave town) is the burning of embassy documents. In the case of the Great War, documents crucial for understanding Serbia's role in the assassination of the archduke did not become available until the 1950s, under Tito.

Finally, while the distinction in history of science between the external and internal history is not a sharp one, there is obviously a great deal to be said for it, analogous to the distinction between the contexts of discovery and justification. The external history involves the pressures of other people upon the reasoning scientist, some of which have a tendency to distort his theoretical thinking in an irrational direction. But the basic cognitive process, despite these impingements, is between the scientist and the nonsocial world of his subject matter; whereas in a psychopathology case study or in political and religious history, the "external world" of the protagonists under study is itself the

actions of other people, with all the complications, murkiness, and defensiveness that implies. Nature, after all, is not trying to make a Dutch book against us (Glymour, 1980; Shimony, 1955, 1967), but the enemy ambassador (or the patient's mother-in-law) may well be.

STATISTICS AND CAUSALITY

There are two ways that statistics applied to case studies can bear upon the question of causality. In the strong use of case studies, the investigator infers certain causal influences to have been at work from his study of the individual case, relying upon a mixture of corroborated theory and commonsensical generalizations (e.g., by and large, people tend to take steps to avoid, or escape from, situations they dislike), and then presents statistics which summarize those case study causal findings. This was the method followed by Freud in his original theory of the etiology of anxiety neurosis versus neurasthenia as due to two kinds of unhealthy sexual practices; and similarly (pre-1897) his theory as to the specific etiologies of hysteria and the obsessional neurosis. He argued from the intensive study of each case (the relationship between the symptoms and the patients' reminiscences under hypnosis or free association) to what events in the patient's current and past life were clearly related to the symptoms of the neurosis. Having reached certain causal conclusions *on the basis of the intensive study of each case*, he then argues (e.g., for hysteria), "If you submit my assertion that the aetiology of hysteria lies in sexual life to the strictest examination, you will find that it is supported by the fact that in some eighteen cases of hysteria I have been able to discover this connection in every single symptom, and, where the circumstances allowed, to confirm it by therapeutic success" (1896/1962, p. 199).

Obviously such a statistical study over "understood" cases will not be persuasive to a skeptic who does not trust the validity of the individual causal inferences *as each was made casewise*. Therefore we have developed a second way of using statistics from the case method which makes it intermediate between the case study method and other methods, wherein we avoid the initial step of imputing causality casewise. Instead, we record the presence or absence, or sometimes the magnitude, of a certain factor that is essentially free of causal imputation, something close to an observable or a summary of observables, categorized in ways that may be suggested by the conjectured causality but not hinging upon its being valid. Then we correlate "output" properties, such as a symptom or neurosis or character trait, and the inferred causal factor. An example of this approach would be showing a statistically heightened incidence of object loss (death or divorce of a parent, death of a sibling or even a much loved pet) in childhood among adult depressives.

These two ways of using statistics with case studies are frequently combined. For instance, we might discover that the onset of an operationally identifiable childhood

depression was close in time to the object loss, and then the incidence of such occurrences might be shown by statistics to be heightened in the life histories of persons who have experienced diagnosable depressions as adults. It is admitted on all sides that both ways of using statistics on case material are problematic, although not for the same reasons.

CLIOMETRICS IN GENERAL HISTORY

Controversies about the value of cliometrics in the study of history continue unresolved to such an extent that the cliometricians have established their own journals and their own societies, as if the cliometricians and other historians have agreed to disagree and even to go their separate ways.[13] I expect these disputes will be interminable unless one makes a distinction between two uses of statistics generally. I have not seen this distinction made by the cliometric disputants, and as a result they are often talking past each other. The distinction was set forth by me many years ago (Meehl, 1954/1996, p. 11-14) and, although I do not quite like my old terminology, I have seen no reason to discard the conceptual distinctions. The *discriminative-validating* use of statistics restricts itself to inferring an association between minimally theoretical variables, introducing no "deeper" theoretical constructs than those at a first level of abstraction (such as human trait names), imputing no causal relations between them or deeper level explanation of their correlations. For example, if one is willing to admit the concept "chronic alcoholism" into one's clinical vocabulary (and a reliable set of operational criteria for that exists), and one is willing to admit the ethnic term "Irish" into one's vocabulary, then it turns out that the only two highly valid predictors of chronic alcoholism are (1) having a first-degree male relative, such as a father or brother, who has been diagnosed as

[13]"Cliometrics: The study of historical data by the use of statistical, often, computerized techniques" (*Random House Dictionary of the English Language*, 2nd ed., 1987. New York). The etymology is from Clio, Muse of history, + metrics. It takes a singular verb. My colleagues Alan Shapiro and Roger Stuewer (History of Science and Technology) comment that one might have expected historians of science to take up cliometrics, given their usual familiarity with quantitative methods in the sciences they study. Oddly, this is not the case, historians of science having shown less interest in cliometrics than economic historians and "culture" historians (Tilly, 1984, p. 365). Keeping in mind that different kinds of historiography may differ widely in the applicability of quantitative methods, readers unfamiliar with the cliometric controversy among historians may consult Aydelotte (1971); Aydelotte, Bogue, and Fogel (1972); Barzun (1974); Benson (1972); Bogue (1983); Conrad and Meyer, 1964; Diamond (1980); Erickson (1975); Fitch (1984); Flanigan (1984); Floud (1973, 1984); Fogel (1975); Fogel and Elton (1983); Fogel and Engerman (1974); Hays (1984); Himmelfarb (1987); Hobsbawm (1980); Jarausch (1984); Judt (1979); Kocka (1984); Kousser (1984); Lorwin and Price (1972); Rabb (1983); Rowney and Graham (1969); Schlesinger (1962); Simonton (1990); Stone (1979); Tilly (1984); Wachter, Hammel, and Laslett (1978). In my opinion, social psychologist Simonton's book is the best single place to begin reading about cliometrics (he prefers ⌜historiometry⌝ for what he does). Conceptually clear, mathematically sophisticated, sensitive to the difficulties and dangers, with fascinating data ranging over several behavior domains—if it doesn't convince you that cliometrics is worth looking into, I suspect nothing will.

alcoholic and (2) being of Irish ancestry. Or again, it is a statistical fact that persons who have made suicidal threats or attempts in the past are far more likely to kill themselves than persons who have not.

In the second general use of statistics, a proper subset of the statistical notation variables are interpreted in the embedding text as counting or measuring highly inferential or theoretical states, properties, or events, which inferred theoretical events have causal efficacy in generating the correlations and time series displayed by the observed variables. This I called the *structural-analytic* use of statistics, although today I would prefer the terminology *causal-theoretical* use. The prototype of this kind of statistics is factor analysis, other examples in psychology being the formalism of classical test theory, multidimensional scaling, path analysis, and the more theoretical portions of taxometrics (mathematics of classification, see Meehl, 1992a; Meehl & Golden, 1982). In this use of statistics the inferential problem goes far beyond the statistician's problem of sampling error, whether biased or random. If the numbers characterizing the observations are accepted at face value, there is a transition from them to numerical values attributed to the theoretical entities, and, of course, a transition to the very existence of those inferred theoretical entities. *Example*: Given the observed correlations over a sample of subjects among the subtests of an omnibus intelligence test, I subject this correlation matrix to a factor analysis, on which basis I assign certain factor loadings to the various subtests. I also characterize the inferred psychological factors by theoretical names such as verbal fluency, induction, spatial ability, or whatever. I can solve a set of equations derived on the basis of the factor loadings to infer the latent factor score of an individual from his pattern of scores on the subtests. Thus I say, "Jones has an IQ of 123, and the best estimate of his standard score on the spatial factor is 1.8." That theoretical statement is warranted by a set of statements which are in turn based upon problematic solutions to a mathematical problem (the rotation problem in factor analysis). The interpretative text, in speaking of ⌜spatial factor⌝, obviously goes beyond a mere first-level characterization of subtest tasks. A psychologist who does not like the way I solve the rotation problem as a problem in applied mathematics or who disagrees with my theoretical interpretation of Factor I (based upon my inspection of the subtest factor loadings) need not accept my attribution of the spatial factor score to patient X, although he may have no reservations about how the test was administered or about any of the descriptive statistics including the original subtest correlation matrix. If he distrusts factor analysis generally, or holds a fictionist metatheory, he may reject *any* attribution of causal efficacy of a mental factor "in the person" as unsound reification of what is merely a convenient reference axis in a psychometric space (Meehl, 1991). Even if the factor solution is accepted mathematically and the factor considered "real," the problem of psychological construal of the alleged

factors is a grave one, still unsolved (but cf. Meehl, Lykken, Schofield, & Tellegen, 1971). One reason that the dispute between cliometricians and "traditional documentary historians" remains unresolved is that the disputants talk past one another about these two uses of statistics, cliometricians emphasizing the indispensability of the discriminative-validating use of statistics, anti-cliometricians stressing the problematic character of the causal-theoretical (previously termed structural-analytic) use. *Traditional historians ought to concede the unavoidability of the first use of statistics, and cliometricians ought to concede the problematic character of the second use*.

Making this distinction, my position is that one should make statistical summaries of the results of intensive case studies of scientific episodes, carefully distinguishing between those in which the statistics summarize *causal inferences made intrastudy* from those in which causality is inferred from the pattern of correlations over studies, without prejudging which of these approaches is better or when. Second, I shall argue that we can distinguish between the discriminative-validating question whether certain kinds of statistical relationships obtain among properties of theories (as distinguished from the behavior of scientists in appraising them) and the other (more interesting) issue of causal-theoretical use, which makes claims about the relation between theoretical properties and inferred verisimilitude.

Looking upon case studies as context of discovery where we are seeking to distill certain guidelines (I repeat, *not rules*) for appraisal of scientific theories, we would supplement the findings of historians in case studies with other sources of conjectures concerning metatheory. Those sources are armchair epistemology, arguments taken from statisticians and probability theorists (e.g., Bayes's Theorem), and metacommentary by working scientists, both metacommentary made informally in the course of theorizing and experimenting and that offered systematically by scientists who have an intrinsic interest in theorizing about "scientific method." In employing this latter source, we do not assume that every scientist is always an accurate introspector or describer of his own scientific behavior but merely that sometimes some of them are. The fact that sometimes scientists are poor at this—it is well known that some scientists have said stupid things about "scientific method" and can be found doing things that their own metatheoretical account does not countenance—would not disturb us because we are operating in the context of discovery. Whatever ideas we get from the traditional armchair epistemologists, statisticians, introspective scientists, historians of science, and logicians are all candidates for cliometric empirical study.

Given a proposed guideline, how do we study it? We study it by collecting cases. Some of these individual cases may be crucial, functioning as falsifiers of a

metatheoretical *rule* of universal form. But since it already appears from the available case studies that there are hardly any such valid rules (Feyerabend would say literally no such), I anticipate that few case studies could function in this crucial way and that some would have greater evidentiary weight than others.[14]

Since the modest claim of a metatheoretical guideline is one of over-all average expectable success, of a recommendation whose acceptance tends to be advantageous more often than not, a system for random sampling of the research literature of a science is imperative. This suggestion does not contradict the habit of historians of science to look intensively at episodes which played a major role or which involved scientists of special eminence. All I am arguing is that if a metaprinciple purports to be a guideline that it is, by and large, statistically advantageous to follow, *that is an intrinsically statistical claim*. It is hard to defend such a claim without computing statistics based upon random or representative sampling of what took place historically. When one has collected such statistics, two questions—both empirical—can be asked about these summary results. First, what do scientists in fact tend to do with regard to a certain guideline? Do they follow it always, usually, seldom, never? Second, does following the guideline tend to work or not?

It is imperative to formulate this latter as *tend*. The scientist is like the gambler, the businessman, the physician, or the psychotherapist. He knows in advance that no matter what policy, guideline, or rule of thumb he follows, and no matter how clever he is in implementing it, he is sampling from the vast universe of facts and he is working much of the time in the semidarkness of more or less ignorance. Hence all guidelines are inherently stochastic, both as to their descriptive and prescriptive roles. That being so, what does a "successful deviation" prove? I think it proves very little, since the guideline is not formulated as a universal rule. We would need to know more details about a particular episode, which it is the traditional historian's job to elucidate. We should also remember that a success in deviating from a recommended guideline can be a matter of sheer luck. Suppose a scientist, biased by personal identification with his PhD advisor's favorite theory, relies on a minority of positive unreplicated studies despite a

[14]"Don't leave out observed facts or make them up" would probably come as close to a rule of empirical science as anything can. Faking data is the unpardonable scientific sin, being incorrigible by the standard, easily accessible means (e.g., recalculation of statistics, detecting a formal fallacy in a derivation, pointing to inferential flaws in interpretive text, discerning inconsistencies between a definition and application). Yet even this "rule" is fuzzy at the edges, and sometimes profitably violated. For example, Millikan in his classic electron experiment ignored several "poor" readings —as shown by his laboratory notebooks—despite alleging in the published work that all were included (Millikan, 1917). His final numerical value for *e* is now considered remarkably accurate, closer than it would have been without the deletions. Whole treatises exist on the theory of omitting "outliers"; and Fisher gives procedures for filling in (i.e., *making up*) "missing values."

preponderance of replicated ones adverse to the theory. This is "irrational" (= poor betting odds, counter to the actuarially successful principle). But perhaps the statistical sampling (by researchers) from the huge possible fact collective has been, through no one's fault, a bad sample of experimental situations (Meehl, 1990b). In the social sciences it can also be due to bad (random) sampling of *organisms* despite representative coverage of *situations*. Or perhaps the scientist is relying on a countervailing guideline that we have not as yet distilled out of our cliometric studies. It could be that one scientist is simply better at choosing experiments, in analogy to trait concepts like "accident prone" and "creative" in industrial psychology. It is obvious from the presently available (nonsystematically sampled) case material that there are occasions in which two plausible guidelines, maybe even two accepted by the same individual scientist, point in opposite directions with respect to a theory being appraised. The question whether a certain guideline, which itself has predictive validity as to a theory's ultimate fate, should countervail another one which has also been shown to have predictive validity cannot be answered unless we have statistics on the same collection of theories. Contemplating the list of candidate principles in the inside cover of Donovan, *et al*. (1988), one sees immediately that most pairs of those principles neither entail nor contradict each other in the sense that a particular scientific episode or problem-situation *could not* be oppositely subsumed by them. Thus countervailing preference rules or a composite index (perhaps with differential weights) is needed.

Here the metatheorist may get help from the psychologist who is accustomed to choosing among various psychometric models in selecting students, employees, or military personnel, where the choice of model involves the statistics of the particular predictive problem, taken together with the relevant utilities and disutilities. In industrial and clinical psychology we speak of a *compensatory model* (also called a *regressive model* because it appears in regression equations), in which high scores are permitted to countervail low scores. The potency of this countervailing effect is represented by the relative weights in the standardized regression equation or discriminant function. Then we have what is known as the *successive hurdles* (*conjunctive*) model, as when we say, "In order to be admitted to law school, an applicant must have a grade point average > K, have gone to an undergraduate college rated > L, and have a score > M on the law school aptitude test." Here a high grade point average cannot make up for an applicant's deficiency in having attended a third-rate undergraduate college. Then we have *disjunctive* models, much less common for obvious reasons, where we say a score $x >$ K or a score $y >$ L suffices for admission. Finally, probably best in many situations, is a *mixed model*, where setting up successive hurdles at reasonable levels still leaves us with a larger proportion of applicants than we can take. We set up a rule:

If $x_1 > k_1$, $x_2 > k_2$, … consider the applicant for admission, but then apply linear equation $y = b_1x_1 + b_2x_2 + b_3x_3$ … and admit him if $y >$ K.

One can confidently anticipate that the theory-appraising scientist, as well as the metatheorist proceeding in anecdotal fashion, will often decide nonoptimally if by ⌜optimal⌝ we mean the best mathematical combination of appraisal variables. I think this can be safely said on the basis of the sizable and varied body of research in social, cognitive, and clinical psychology. It is of course tempting for one to conclude that scientists by and large proceed with high cognitive efficacy because science is so much more successful than other purportedly cognitive enterprises that do not seem to get anywhere in settling their problems. But that inference is a mistake.[15] That physiologists and astronomers do better at answering their questions than metaphysicians, ethicists, literary critics, theologians, politicians, or journalists do at theirs does not tell us how well, in terms of an optimal or ideal, the scientists do. Humans attempting to combine quantitative information for a certain predictive or decisional purpose do rather badly, partly because they do not compute statistics very accurately "in their head" compared to the statistician using a formula, and therefore *assign nonoptimal weights* in their subjective regression equations. Second, they do less well than is possible from the information because they *apply those weights inconsistently* (as the psychometrician would say, "unreliably"). This important fact is illustrated by the Goldberg Paradox, named after the psychologist who discovered it in an illuminating and ingenious experiment reanalyzing some of my data on clinical prediction. Goldberg (1970) showed that, if clinicians attempting to make a diagnostic distinction between neurosis and psychosis from the Minnesota Multiphasic Personality Inventory are asked to sort patients' profiles on an eleven-step scale, they do not do as well as a simple unweighted linear composite of certain of the MMPI scores. If one sets up a regression equation whose weights optimize prediction of the clinicians' ratings of the patients (rather than taking as criterion the "correct answer," the patients' diagnoses) and then applies this strangely derived equation to the usual predictive task, *it does better than the clinicians' judgments on which it was derived*. This was true both for the individual clinicians and their own equations and for the clinicians as a group. This counterintuitive finding, which astonished most clinical psychologists when it was published, is—once somebody as clever as Goldberg thought of doing it—quite easily explained. It arises from the fact that the clinician does not apply his own (admittedly suboptimal, but still somewhat valid) weights consistently. Thus the inadequacy of the clinician's prediction of the true diagnosis arises partly from his not employing the best weighting scheme and partly from

[15]The best general treatment is by Faust (1984).

his applying his own weighting scheme unreliably. When we employ Goldberg's equation for "predicting the clinicians" to the real task of predicting the predictand, the mathematics that stochastically models the predictor therefore does better than the predictor himself.[16]

QUANTITATIVE OVER QUALITATIVE APPROACH

I said above that I do not share the enthusiasm of some of the younger metatheorists for the potential contribution of the *empirical* cognitive sciences (cognitive psychology, economics, sociology of knowledge, political theory). I expect any substantial cognitive science contributions to be (a) from the more formal of them, such as decision theory and statistics, and (b) negative, alerting us to the deficiencies of the human mind as an information processor and to social pressures against rationality. But as a psychologist I do have one positive suggestion which I believe would yield not minor improvement but marked benefit to empirical metatheory, namely, the replacement of qualitative by quantitative concepts in most real-life contexts. It is natural for American trained psychologists to think in terms of factors or dimensions—matters of degree rather than kind—because we learned the hard way that putative categories, types, or taxa usually turn out, on careful inspection, to be regions on dimensions. For example, in the realm of "normal range" personality description it is doubtful that there are any true personality types; instead one finds clumps or clusters of persons who are located in a certain interval of a psychometric factor (continuum) or, when we deal with many dimensions, are located within a certain volume of the descriptor hyperspace. I do not mean to deny the utility—and the theoretical necessity—of defining genuine taxa in certain domains of psychology, especially psychopathology, where research in taxometrics has been one of my major concerns in recent years (Meehl, 1992a; Meehl & Golden, 1982). It has traditionally seemed natural, going back to Aristotle, for philosophers and logicians to think in terms of categories and the qualitative predicates that define them, and I do not complain of this. But in formulating metatheory based upon the facts of scientific development, this "class" or "property" predilection gets us into trouble which a quantitative (dimensional) approach would avoid. My reading in this area convinces me that the examples are ubiquitous and could easily be shown with a content analysis to be literally in the hundreds. I confine myself to two examples from the excellent collection of case studies in Donovan, *et al*. (1988).

In his case study of the vortex theory of motion, Baigrie examines the thesis about assumption GA2.3 (see Donovan, *et al*., 1988, endpages) that "scientists often refuse

[16]A formal treatment of the general conditions for Goldberg "bootstrapsing" of human judgments, together with a list of 15 empirical studies (so far, it always works!) can be found in Camerer (1981).

to change their guiding assumptions," and he points out that this has been interpreted in a Kuhnian light, signifying that scientists are often dogmatic about their theoretical commitments. He goes on to say that this interpretation "has occasioned a great deal of theoretical activity on the part of philosophers, since it is not immediately apparent how dogmatic behavior on the part of scientists can be reconciled with our supposition that science is the apex of rationality" (Donovan, *et al*., 1988, p. 85). Despite the semantic help offered by the "often" in the original thesis, Baigrie seems to treat this empirical generalization as though it were not a matter of some (unspecified, but non-negligible) frequency of occurrence but rather a nomological. I say this because absent that all or none way of viewing it (despite the term ⌜often⌝) there is no contradiction of the sort he suggests. Suppose we expand the principle in a way that I assume the editors who made the list would find unobjectionable as a more precise description of the empirical situation, thus: GA2.3′ "*Some* scientists are *sometimes* dogmatic about *some* of their guiding assumptions." We don't state whether the "some" means a few, a sizable minority, a majority, or the overwhelming preponderance. But we are careful not to say that *all* scientists are *always* dogmatic about refusing to change *any* of their guiding assumptions, an absurd thesis which surely nobody (including Kuhn or Feyerabend) has ever maintained.

Now how does the revised statement GA2.3″ conflict with our supposition that science is the "apex of rationality"? Obviously it does not conflict with it, even a little bit. When we say that something is the apex of something we do not mean that it is perfect, or infinite, or = 1 on a scale of 0 to 1. All we mean by ⌜apex⌝ is a maximum, that is, higher than anything else around. To say that science is the apex of rationality means that scientists are more rational, collectively and in the long run, than are preachers, journalists, politicians, and (perhaps) scholars in some other disciplines such as literary criticism. GA2.3′ doesn't even tell us whether scientists are rational *most* of the time. The level of rationality in other domains is so low (does anyone dispute this truism about the human condition, especially in matters like politics or religion?) that scientists might be rational only forty percent of the time and science could still be the apex of rationality.

For another example, somewhat subtler but still to the point, I take Finocchiaro's case study of Galileo's Copernicanism where he quotes Galileo's letter to Kepler to the effect that Galileo has not dared to publish "because Copernicus, although he earned immortal fame with some, nevertheless became the target of ridicule and scorn with innumerable others (such is the number of fools)" (Donovan, *et al*., 1988, p. 54). This passage is taken to show that Galileo did not think he had extremely strong arguments because "he obviously does not think they are conclusive or even strong

enough to convince someone who, unlike Kepler, is not already favorably inclined." This may in fact be a correct statement about Galileo's state of mind, but it surely does not follow from the quoted passage. Galileo might think that he had very convincing arguments for Copernicus's position but prudently refrained from publishing because of the number of fools. When you classify a large number of people as being fools, part of what you have in mind is that they are too stupid, ignorant, or dogmatic to be convinced even by good arguments. I find nothing in the quotation to suggest that Galileo means that one must be "already favorably inclined." Here again the problem is that we should have to formulate the strength of arguments as a matter of degree, the resistance of "fools" as a matter of degree, and the number of persons whose mental habits and abilities are foolish as a percentage, before we could know how much this tells us about Galileo's personal doubts regarding the evidence favoring Copernicus.

The psychologist's rational expectation of finding quantitative dimensions where others would speak of categories must not be turned into a dogma that there *are* no real classes, types, or taxa in the world. Quantitative methods are useful in detecting latent taxa and sorting individuals into them. Whether a putative category is real, or merely an arbitrary rubric used for convenience to locate entities in a roughly demarcated region of a descriptor hyperspace, is a matter for empirical investigation employing suitable taxometric methods (Meehl, 1979, 1992a; Meehl & Golden, 1982). The continuing dispute as to whether Kuhn's scientific revolutions differ from the puzzle solving of normal science "in degree" (bigger, harder puzzles!) or "as a kind" should be resolved by taxometric analysis of the statistical relations among the indicators Kuhn has listed.

VERISIMILITUDE AND THEORY PROPERTIES

I would like the development that follows, with some specific suggestions for quantitative appraisal methods, to be as free as possible of metaphysical or epistemological commitments. I should lay my cards on the table and say that I am myself a scientific realist rather than an instrumentalist, fictionist, or pragmatist; but, although I will discuss them in a realist framework, I do not believe that the quantitative suggestions I offer hinge upon that. My suggestions could be acceptable to one who opted for, say, Charles Sanders Peirce's formulation: "The opinion which is fated to be ultimately agreed to by all who investigate, is what we mean by the truth, and the object represented in this opinion is the real" (1878/1986, p. 273). Since I am myself a scientific realist, I will formulate the relationship between the concept of theoretical success and the quantitative indicators of a theory's performance in terms of verisimilitude. While it is admitted on all sides [including by Popper (1962, especially pp. 215-247; 1972; 1983,

pp. xxxv-xxxvii; Schilpp, 1974, pp. 1100-1114)] that at present no satisfactory definition of verisimilitude has been constructed, I believe the concept is indispensable to the scientist whether he has ever heard of Popper or not. Scientific theories are like newspaper accounts or "historical novels" in that they can vary from zero verisimilitude, totally made up as a piece of fiction having no factual reality, to a liberal mixing of truth and falsehood, to a long story in which everything is completely accurate except that, let us say, one person's middle initial is erroneous. It is obvious that the kinetic theory of heat has much higher verisimilitude than the caloric theory, that the van der Waals correction has greater truth likeness than the uncorrected PV = RT, and so on. The term means what the Latin etymology says, "truth likeness" (nearness to truth, better approximation, closer to the objective facts, more accurate model). The clearest example showing it is *somehow* a matter of degree is the case of two theories identical in their formal structure and operational ties, asserting the same mathematical functions, but the parameters of one are numerically closer to the correct values. Speaking as a working scientist who wants to work at better theories rather than poorer ones and who takes truth as a regulative ideal, my rejoinder to my philosopher friends when they object to my mentioning verisimilitude is that if efforts to define it with the familiar tools of the logician (as Popper and others have attempted, e.g., in terms of a consequence class of propositions) don't work, they should go back to the drawing board and approach the problem in different ways until they come up with something that does work. [For the logicians' efforts at explicating verisimilitude, see, e.g., Goldstick and O'Neill (1988), Hilpinen (1976), Kelly and Glymour (1989), Miller (1972), Newton-Smith (1981), Niiniluoto (1984, 1987, 1991), Oddie (1986, 1990), Popper (1962, Chap. 10 and *Addenda*, 1972, Chapters 2, 3, and 9, 1976, 1983), Tichy´ (1978), Tuomela (1978), and a brief summary of the difficulties in O'Hear (1980, pp. 47-56).] I have myself made some tentative gropings in that direction (see Meehl, 1990a, 1990b) which I will not detail here but only summarize.

Briefly, Table 2 lists aspects of similitude between two theories. Since T_{OJ} ("Omniscient Jones's theory") is literally true, the verisimilitude of a theory T is its similitude to T_{OJ}. Consider a postulate P_i of *T*. It can "pass" or "fail" at each level I-X, and these verisimilitude (*V*) levels are (nearly) Guttman scalable—a postulate can hardly pass at any *V*-level higher than the first one it fails. Each postulate gets a "score" equal to the number of levels (1-10) it passes. One crude index of *T*'s verisimilitude would be the mean (or percent) of levels passed by all its postulates. Thus, a 9-postulate theory could score from zero to 1.00 in verisimilitude, standardized as $V\text{-score} = m/(9 \times 10)$. I find this simple postulate batting average upsets philosophers (*and* scientists), mainly because one has a strong intuition that the levels are of unequal importance. I share this intuition, but the trouble is that (a) people's intuitions as to relative importance disagree (see Appendix

TABLE 2
PROGRESSIVELY STRONGER SPECIFICATIONS IN COMPARING TWO THEORIES (SIMILITUDE)*

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

*See (Meehl, 1990b, p. 17; an earlier version was published in Meehl, 1990a).

2, pp. 463-467) and (b) intuition in such complex matters, while deserving our respectful attention, is not a safe guide as we proceed to objectify and quantify. *The improvement of a crude index in metatheory must await empirical study of its statistical properties, in reflective equilibrium with whatever logical, semantic, epistemological, and mathematical analyses can shed light on its successes and failures*. We have left the Vienna Circle armchair (or, I prefer to say, have *added* history books to our reading matter while ensconced in it); nevertheless, we plan to *reflect* on the actuarial facts, as we do in any empirical discipline. To further assuage the reader's cognitive anxiety, I point out that the difference between two weighting schemes is not worth fretting over at an early stage of quantification (see statistical references in Meehl, 1990b, p. 19, bottom paragraph). As philosopher Herbert Feigl used to admonish social scientists, "Why cut butter with a razor?" In the present instance, the psychometric theorems about convergence of weighting systems are especially reassuring, because the Guttman scalability of levels passed forces a high average correlation between number of postulates passing any level and any other level; and these correlations play a major role in the several formulas for estimating inter-weighting agreement. For the benefit of fictionists and instrumentalists, I have also mentioned the possibility of weighting verisimilitude levels by maximizing their composite correlation with a best composite weighting of indexes of theory performance, the industrial psychologist's *canonical correlation* (Meehl, 1990b, pp. 18-19). I do not, of course, offer this list of specifications except as exemplary, since the components of verisimilitude will differ among sciences.

The theory properties can be divided into internal or intrinsic properties (formal + conceptual) and external or extrinsic (empirical + psychosocial). Conventionally philosophers of science are not supposed to pay attention to the psychosocial, which do not belong to Reichenbach's context of justification. While I admit Reichenbach's (1938)

distinction and consider it obscurantist to try to fuzz it up any more than it needs to be, I do not wish to prejudge the extent to which psychosocial properties of theories are to be included in the predictor list.

As in constructing a personality test from a pool of candidate items or a predictor of college achievement from a battery of test scores and life history facts, we do not despair because no gold standard criterion is available. Item analysis to develop an inventory scale for detecting schizophrenia or measuring depth of depression only *begins* with a formal psychiatric diagnosis or perhaps a rough quantitative rating scale by clinicians on the relevant dimensions. It is *never* assumed that such clinical judgments are infallible, either in rating the individual traits that belong to a diagnostic cluster nor in applying the diagnostic category itself. It is a short exercise in high school algebra to show that one can, by statistical analysis of items against a highly fallible criterion, construct an MMPI scale that is superior to the criterion itself. This was understood already by Binet and Simon when they built the first effective intelligence test using chronological age, age grade placement, and teachers' judgments as the "criterion" of intelligence which a good test item should *tend* to follow.

This process was characterized by Cronbach and Meehl (1955) as the "bootstraps effect," and, while it has a special methodological interest to the psychometrician (chiefly as an antidote to simplistic ideas of validation), it is found in all fields of science. We trust a thermometer more than we do the human hand in assessing how hot the soup is, but the first stage criterion was the human hand. Starting with chipped flint tools in the Pleistocene age, we now have methods of polishing surfaces "flat" within a few molecules thick. Suppose an epidemiologist had blindly tested a group of neurological patients diagnosed as general paresis in 1900 (when the luetic ideology was widely held but still not proved and not accepted by all physicians), giving a big battery of miscellaneous biochemical tests (Schick, Mantoux, Wasserman, spinal fluid colloidal gold test, etc.). He would surely have discovered that the patients called paretic differed *statistically* from the patients labeled with some other neurological disorder, including those with dementing and other psychiatric changes, the paretics having a positive spinal Wasserman and a first zone colloidal gold pattern more frequently than the others. The Schick (diphtheria) and the Mantoux (tuberculosis) would not "work" and would be dropped from the candidate list. But it may well be, looking back today, that the accuracy of diagnosis of dementia paralytica in 1900 was only 80% valid, so that if the joint sign (positive Wassermanfirst zone colloidal gold curve) had been taken as a new diagnostic indicator for clinical use, it would have appeared to be only 80% correct, whereas in reality, as known to Omniscient Jones, it was performing almost 100% correctly.

Because the whole procedure is statistical and whatever final rules of thumb the metatheorist arrives at on the basis of his actuarial study of theory properties in relation to "best bet" verisimilitude will only be probabilistic, it does no harm if some theories ensconced in the textbooks and universally believed today, used as best available criterion cases, are subsequently found to be false after all. We do rely on the assumption that if a theory has been in the textbooks for, say, 50 years or even 25 years, and is generally not referred to any longer in the textbooks as a "theory" but as an established fact (the scientist often writes this way, although the philosopher knows it is always going to be a theory no matter how well corroborated), such a theory is *highly unlikely* to be dislodged in the next 50 or 100 years, and most such will probably not be dislodged before the sun burns out.

When a study based on a sampling of history of science episodes has been concluded within a domain, the metatheorist has the materials for rational reconstruction, for explaining why science works better than other alleged cognitive enterprises. We also have, as a result of the actuarial generalizations, a set of guidelines for the use of the working scientist, even one not intrinsically interested in metatheory, in his unavoidable efforts at theory assessment.

Thus the basic procedure would be to define a domain, sufficiently broad to lead to interesting generalizations but sufficiently narrow to allow for possibly marked differences over domains (it would be surprising if chemistry and personality theory gave the same statistical weights, a matter we would not wish to prejudge), going back far enough in history so that we have a set of criterion theories that we take to have very high verisimilitude, our best approximation to a gold standard criterion. We must choose theories well enough along at an early period that we can formulate the postulates (whether or not theorists at the time did so explicitly) and theories that also had a long enough empirical life to enable us to do statistics on the factual side, the empirical track record, among the properties. Or we might consider each theory at its experimental half-life, counting to half the total number of experiments performed before it was either abandoned by everyone as dead or enshrined in the textbooks as being the clear truth of the matter. Here again, one should not be distressed by the apparent arbitrariness of the choice of such theory "ages," because we rely on the basic scientific principle that *what brings order into the material is the correct way to do it*. The whole set of theory properties are then combined to yield the highest correlation with the *V*-scores of competing theories as well as the gold standard one. The gold standard true theory, which we are for bootstraps purposes treating as if it were T_{OJ}, may be included in the batch. If that bothers anybody as "circular" (which it is not, in any bad sense) we can leave the true theory out and consider only the theory under study and its competitors that

fell by the wayside. We know this is a biased sample in a number of respects even if we began by a random sampling of theories in a domain, because such constraints as how quickly the theory died or how financially expensive it was to test some of its predictions and the like will eliminate some theories from the list of failed competitors because data are insufficient. *The important thing is not to become irrationally perfectionist in the context of discovery at this stage.*

The intrinsic theory properties I divide into *formal* (features of the logical and mathematical structure) and *conceptual* (properties revealed in the embedding text that interprets the formalism). I emphasize a distinction not usually made between two subdivisions of the embedding text, one the *operational text* that coordinates a proper subset of the theoretical concepts to the observation language, and the other the *interpretive text* which characterizes the theoretical entities in various ways (Meehl 1990a, p. 109; 1990b, pp. 2-5). The most important conceptual category is the simple one of what kinds of entities are postulated, and then what are the relationships among them. The three main relationships that can exist among theoretical entities, so far as I know, are *efficient causal*, *structural-compositional*, and *developmental*.

The external or extrinsic properties of theories are the factual ("observational track record") and the psychosocial. The latter are questions that we normally put in the context of discovery but there is no reason not to include them for *this* purpose, because, after all, these questions are part of the history of science. Everyone knows that scientists include certain psychosocial properties in their appraisal of theories and in deciding on their own research programs and fund granting agencies rely on them very heavily. So do some editors, and not—I think unwisely—some others (e.g., teachers, students, researchers, practitioners). However, considering psychosocial properties such as "Was this theory invented by Einstein?" or "Is this theory highly regarded by clinical practitioners?", it is unclear to what extent a *scientific article* assessing the state of a theory on present evidence should pay attention to them. My own tentative view is that the situation is rather like that in the law courts, where certain kinds of evidence are inadmissible in a criminal trial on the grounds of being "prejudicial," despite the fact that everybody knows that they would be highly relevant if looked upon in a purely statistical fashion. For example, a defendant's previous criminal record is highly evidentiary from the rational point of view, perhaps statistically more predictive than, say, identification by a doubtful eyewitness, but we exclude it nevertheless. Similarly, ethnicity is strongly predictive of certain behavioral dispositions, but we require that the individual's dispositions be *sampled*, however unreliably, rather than relying on race. The probabilities that somebody is guilty of drunken driving if he is of Irish ancestry are almost 20 times as

high as if he is Jewish, but nobody suggests that this be admissible in evidence at a criminal trial. This is a complex question, and beyond the scope of this paper. I suspect that one reason the older generation of scientists and philosophers becomes nervous about the "empirical metatheory" approach is that they fear that these kinds of previously excluded considerations would begin creeping in. I sympathize with this worry, because I believe that post-Galilean science is better than medieval science partly because arguments from authority are not countenanced officially by the rules of the scientific game.

Among the formal (logical and mathematical) properties of theories that should be included in a candidate list of predictors of verisimilitude are the following, which I will simply list without expansion or defense: To what extent are mathematical functions relating theoretical variables to each other, or to input or output observational variables, simply *postulated* versus *derived* from more basic and sometimes qualitative postulates? For example, Guthrie's decelerated learning function follows as a consequence of his single cue one-trial yes-or-no connectionism, whereas Hull's decelerated function for habit strength as a function of the number of reinforcements is a basic postulate in his system. What is the ratio of the number of entities to the number of postulates (this being an aspect of "simplicity")? What is the ratio of the number of core postulates to the number of peripheral postulates (as defined in Meehl 1990a, 1990b)? What is the proportion of the observational well-formed formulas (*wff*s)[17] requiring peripheral postulates rather than flowing from the core postulates alone? What is the ratio of the number of theoretical predicates and functors to the number in the experimental domain? [Conjecture: Theories are scientifically inadmissible if this ratio > 1, although the number of theoretical *entities* may exceed the number of observational macro-objects (e.g., electrons versus chairs and cannonballs).] Given a definition of postulate pervasity in terms of the number of experimental *wff*s in whose derivation chain a postulate appears essentially, we can calculate various kinds of indexes of pervasity for a theory. Are there multiple derivational paths involving slightly overlapping or perhaps ideally even nonoverlapping postulates terminating in the same observational *wff*? What is the ratio of postulates to observational *wff*s? If we compare core versus peripheral postulates, to what extent are the parameters occurring in their functions derivable rather than requiring adjustment by empirical curve fitting? What is the average number of postulates in which the same theoretical construct occurs? How does this number compare between core and

[17]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.

peripheral postulates? How many *V*-levels are covered (ignoring whether they are in accord with the accepted theory)? And then there is some kind of notion of "theoretical interknitting," which surely has several distinguishable (although perhaps correlated) aspects. For instance, in a previous paper (Meehl, 1990b, p. 6-7) in stating assumptions underlying my derivation of a high correlation between experimental success as crudely defined by the number of experiments that "come out right" and verisimilitude defined as how many postulates have been altered from T_{OJ}, I forbid input-output theories in which every input-output relation is independent of all the others so that there are no interconnections between them, calling such relations *isolates*. Scientists do not usually countenance theories in a variegated domain which consist of nothing but a heap of isolates. But permissible theories would differ with respect to how far they are from that forbidden condition, and this dimension of theories would involve complex relations between postulates in terms of shared theoretical terms. In order for any scientific theory to be interesting and fertile, it must of course have shared theoretical terms between postulates, otherwise the derivation chain would constitute a 4-terms fallacy. But the "network richness" of one theory may be very much greater than another even though neither of them consists of merely a heap of unconnected input-output relations. My conjecture is that different aspects of internal interknitting will turn out to be among the most important differences between good and poor theories, especially between "strong" and "weak" theories; but I have at present only intuition for thinking that. Some aspects of interknitting will be quite complex, as the following two suggestions illustrate. *Example*: Given a *k*-term theoretical vocabulary, if *m* of the $\binom{k}{2}$ term pairs $\theta_i\theta_j$ occur linked in a derivable *wff* (testable or not), what is the value of $m/\binom{k}{2}$? *Example*: List all theoretical terms θ_{op} linked directly to an observational predicate ("operationally defined"). Order these terms θ_{op} by frequency of different experimental *wff*s they appear in. Call those with high frequency "observational pivot concepts." Define experimental subdomains by the presence of a pivot concept (one or more). Then there are n_d observational subdomains. For each *non*operational concept θ_{nop}, in what proportion of the n_d subdomains does θ_{nop} occur in a derivation chain? What is the mean and standard deviation of these proportions over the set of θ_{nop} terms?

WEIGHTING COMPONENTS OF VERISIMILITUDE

We come now to the problem of assigning weights to these components to form an index. This does not present a difficult problem for *properties* of theories, given the scientific realists' aim of using them as a means for inferring verisimilitude, since finding the "best" composite of properties, qualitative and quantitative, is a problem in

mathematical optimizing largely solved by statisticians. The quantified internal and external properties of theories (x_1, x_2, … x_m) are to be combined in a function $\hat{V} = F(x_1, x_2, \ldots x_m)$ such that $R_{V\hat{V}}$ is a maximum.

The problem of assigning weights to distinct components presents itself only on the criterion side, for the ten levels of verisimilitude presented in Table 2, which I will call *V-levels*. In our (preferably large) collection of competing theories, we have a criterion set of long-time textbook-accepted theories treated as a verisimilitude criterion. Then among the competitor theories, which of course could range from only one competitor to numerous competitors of a given criterion theory, we have varying degrees of verisimilitude. Each competitor is compared with the criterion theory, the candidate theories' postulates being examined as to their correctness at each of the ten *V*-levels. We may think of each candidate theory as having ten "scores," analogous to the scores an individual subject might get on ten different mental tests in a psychological assessment battery, here consisting of the number of postulates passing each level. I will confine myself now to 10-postulate theories for ease of discussion. (There is *no connection* between there being ten levels and also ten postulates; supposing ten postulates for each theory was done merely for convenience in calculations; having ten *V*-levels, of course, is a result of the ten listed in Table 2.) These scores are quasi-Guttman scalable since it is conceptually impossible for a postulate in a theory to pass level $k + 1$ if it has failed level k. This is important because it helps us dispose of the weights problem. We then have a composite *V*-score for each theory obtained by somehow combining the ten scores (number of postulates passing at each level) that it achieved over the ten *V*-levels. The question arises how should these level scores be weighted?

One could simply add raw scores for all the theories with a given number of postulates, but the psychometrician warns us that this scheme amounts mathematically to assigning weights proportional to the standard deviations. In some circumstances this is harmless, but we should know that we *are* doing it, willy-nilly, if we simply add raw scores. Others would say that, if we don't have any very strong reason for considering some *V*-levels to be more important than others, we should transform the scores at each level into standard scores based on the distribution of theory scores, which is what "equal weighting" means in psychometrics. However, the scientific realist (and nobody else is concerned with verisimilitude) may object that some of these levels are more important in the verisimilitude concept than others. Surely it is more important for a postulate to pass Level I—to postulate the right kind of entities—than it is to be correct to two decimal places in the parameters found in a function at level X? I share the intuition that some levels are more important than others, but it is hardly conceivable that all scientific realists would assign identical rankings to all 10, although they would probably agree that, by and large, those closer to level I are more important than those toward

the end.[18] Intuition is a legitimate help in the context of discovery, and may sometimes be invoked, if strong enough and consensual enough, in the context of justification. But nobody would in a context like this push it as an infallible criterion. Further, what are our intuitions really *about* in considering verisimilitude levels? Are we merely saying that we *care more* about a theory's postulating the right kind of entities than we do about its parameters? Or do we have the notion that there is some deeper underlying unidimensional factor of verisimilitude of which these are in some sense "indicators," as in factor analysis or taxometrics? I have great difficulty in conceiving what sort of dimension that would be, although I do not wish to be dogmatic on the subject.

What we have to deal with here is the old *problem of index numbers*, of how one *conceptually* interprets the question, "What is the optimal composite for an index based upon qualitatively disparate properties or entities?" If an index is intended to be the strongest correlate of something else, then we have the statistician's answer. That is how the various aspects of theory properties are going to be combined once we have the verisimilitude criterion as the variable to be predicted. But that does not solve the weighting of the criterion components.

Lacking an external criterion with reference to which empirically derived weights can be assigned, it might seem that there is a conflict between assigning equal (standard score) weights or going by the realists' intuitions; and if the latter, whose? One need not do an empirical study of realist opinions about the ten levels of theory specification in Table 2 to be quite confident that there would be a nonzero correlation among realists as to their intuitions; but we can be equally sure that the correlation would be imperfect. I myself, concocter of the list, have changed my mind "intuitively" about a couple of orderings over the passage of time. Fortunately this problem is not a serious one because of the well known fact, shown by numerous empirical studies with real data, Monte Carlo investigations, and analytical derivations, that two weighting systems will correlate very highly so long as certain quite general conditions (e.g., positive manifold of the variables) are met. The classic paper is that of Wilks (1938), where it is shown that if we were to reassign a set of βs in a random fashion the expected correlation between two linear composites of a set of variables $x_1, x_2, \ldots x_m$ will be to a good approximation $1 - (1/m)$ (cf. Meehl, 1990b, p. 19, paragraph 3; cf. Richardson, 1941; but see also McCormack, 1956). At this stage of our thinking about cliometric metatheory, that standard finding of classical psychometrics might be sufficient to reassure both social scientists and philosophers.

[18] I leave this erroneous prediction (see Appendix 2, pp. 463-467) as it stood in first draft, a beautiful example of the Faust-Meehl contention that scientists' informal, intuitive, common-sensical metatheoretical habits are not always "objective" in the positivist sense of intersubjectively reliable. My confident meta-prediction here turned out to be utterly worthless.

Despite this reassuring truism of psychometrics, it should be remembered that the increment in predictability produced by a given increment in the Pearson r is not linear in r, and it becomes *very* steeply accelerated when we achieve correlations >.90. Since in social science we almost never achieve correlations of that size with either an external or inferred latent criterion (as in factor analysis), this marked acceleration in predictive power in the region $.90 < r < 1.00$ rarely concerns us. However, when verisimilitude is the concept and a huge population of experiments exist whose outcomes have objective correlations with verisimilitude, we could very well be operating in that steeply accelerated region of validity. For example, I have shown (Meehl, 1990b) that even considering a crude conception of verisimilitude (number of postulates in T_{OJ} from which our theory departs) and an equally crude measure of a single theory performance property (number of experiments that "come out right," not attempting to quantify the closeness or the antecedent risk), one achieves a correlation between verisimilitude and empirical track record in the .90s. It is therefore practically certain that a more sensitive measure of verisimilitude based upon passage of the ten specification levels by the theories' postulates and upon multiple measures of formal and factual theory properties—even a combination of the four indexes suggested below and paying attention to nothing else—could be expected to yield $r \simeq 1.00$. Although this might hold only for the humungous number of actual and possible experiments in a given scientific theory domain that are performed before the sun burns out, I believe we would be very close to it—correlations of .98 and better—when the number of experiments is in the hundreds or thousands. For that reason, *nonoptimal weightings of the verisimilitude levels could reduce this composite predictability from theory properties by an amount large enough to be worth eliminating if possible*. McCormack (1956) has shown that despite the well-known fact of high correlations between two differently weighted linear composites, their *validities* may differ appreciably, especially when the latter are in the very high range.

A sophisticated reader of this paper in draft asked why I expend effort on the weighting problem instead of simply relying on the well known psychometric finding "Weights hardly matter." Let me show numerically why they *do* matter—quite a lot—in this special context. A metatheorist, contemplating empirical research premised on the Faust-Meehl Strong Actuarial Thesis, may be dismayed by the prospect of sampling hundreds of theories from the history of science to get correlations between various theory properties and their long-run verisimilitude. We could reassure such a potential convert by showing how an accurate estimate of the correlation coefficients is attainable by randomly sampling a small number of theories, provided that the true correlations are sufficiently close to 1.00. This is because the dependence on r of Fisher's z_r-transformation $z_r = \frac{1}{2} \log_e [(1 + r)/(1 - r)]$ in the high correlation region is markedly

nonlinear, where the random sampling distribution of r itself is extremely skewed and has very low variance as r approaches 1.00. Suppose the true correlation between a verisimilitude index and a theory property index (over the population of hundreds of theories and thousands of mini-theories) is $r = .995$ [$z_r = 2.9945$], which I consider easily possible, for reasons given above. We sample only 19 theories from a domain (e.g., biochemistry, animal learning, psychopathology). The $SE_z = (N - 3)^{-1/2} = .25$; a 95% confidence belt takes us down to $z_r = 2.3495$ [$r = .982$]; so we could suffer a downward random sampling deviation that would reduce the shared variance by .026, or 2.6% of its true value. Now consider the case of a true correlation, $r = .950$ [$z_r = .8318$], which conventional social science habits would view as "just about as good [as $r = .995$]." The lower edge of the 99% confidence interval is at $z_r = 1.1868$, corresponding to $r = .83$, a shrinkage in shared variance of .214, or 23.6% of the correct value. So the difference between true rs of .995 and .950 means almost an order of magnitude in percent variance accounted for.

While the correlation between two composites based upon different weightings of the same set of variables is conceptually different from a reliability, its statistical effect upon validity is comparable to a reliability; that is, the obtained validity is attenuated by the imperfect correlation between two composites. Intuition suggests to me that the analogy is sufficiently close that the old attenuation correction, dividing by the geometric mean of reliabilities, would be applicable, but I do not offer a proof of this. However, on Burt's (1950) assumption of *randomly* assigned weights, the correlation between composites is attenuated precisely as in unreliability [= random "error" component], so the usual numerical correlation applies. If we were to treat the Wilks $r = 1 - 1/m$ as "good enough," a nonoptimal weighting of a verisimilitude index (Vindex, see Appendix 1, pp. 456-462) would reduce a true verisimilitude-property correlation of .995 to $(.995)(.90)^{1/2} = .81$, quite unsatisfactory if we have only 19 theories to examine.

One solution to this problem (Meehl, 1990b) is the canonical correlation employed by industrial psychologists. If an employer literally has no strong intuitions about the relative "importance" to him of the various components of a job analysis, then the psychometrician, taking him at his word, may assign weights to components of the test battery *and* weights to the aspects of job performance on the criterion side, such that the test composite and the job performance composite will correlate maximally. Unless the result of that procedure came up with V-level weights grossly incompatible with the realists' intuitions, it would seem a nice solution to the problem. Without a rigorous demonstration, I think it is pretty clear that it would not do gross violence to intuitive weights, because those V-levels which almost all of us would weight heavily—such as specifying the right kind of entities, specifying correctly which ones go up and down

together, and whether the function is linear or decelerated—will surely be more potent in their influence on the empirical track record, hence the canonical correlation will weight them more heavily than those toward the other end of the list.

As I have briefly discussed elsewhere (Meehl, 1990b, p. 18; 1992b) different kinds of composite scores or indexes in the social sciences have different sorts of theoretical meaning for the realist. One may think of the composite of an omnibus intelligence test, even when not optimally weighted in terms of subtest loadings on the big general factor, as attempting to estimate some ubiquitous property of brain function. If the rotation problem were solved by maximizing heritability of that first factor, one might daringly—but not crazily—conceive of the heritable component of *g* as literally a count of the number of "bright" versus "dull" polygenes (Meehl, 1991). Moving to the consumer price index (C.P.I.) of the economist, despite its great value for theoretical and technological purposes, we are not sure whether with it we have in mind some really existing factor or dimension analogous to the brain function basis of general intelligence. The index combines the prices of apples and oranges, and the problem of how to do that, when the proportions of a family's expenditures on different kinds of goods and services are known to vary widely with the family's income level, presents a difficult conceptual problem for the welfare economist. Some have held that unless we are willing to postulate something about a cardinal interpersonal utility metric, the problem is insoluble, and no more "reality status" or "validity" can be assigned to one index number than another. Consider next an index of social class, such as the Hollingshead index or the Sims Score Card. Despite the great importance of the SES (socioeconomic status) variable to psychologists, sociologists, and political scientists, I have never met a social scientist who thought there was some sort of *real underlying dimension*, SES, of which the things we can count are merely "fallible phenotypic indicators." In the original study by Hollingshead and Redlich (1958, pp. 387-397) their three index components, taken together with two case readers' ratings based upon detailed study of several hundred families, satisfy the tetrad difference equations (my calculations) and hence can be "explained" by a single common factor. Somehow this reassuring statistical result does not lead us to infer a latent causal entity SES that has the kind of physical reality *perhaps* possessed by utility or satisfactions in the C.P.I. and plausibly possessed by polygene count as the heritable component of *g*.

Now our intuitions about the importance of the different theory *V*-levels differ not only as to weights we would assign (although, I repeat, I think these differences would not be large),[19] but one's intuition as to whether there *exists* some

[19]See Footnote 18!

true valid meaning of verisimilitude, somehow underlying or pervading this ten variable composite, are quite weak. I myself doubt that there is any such thing, so that for me the "best weighting" cannot have the kind of meaning of "best" it might have when speaking of the subtests of the WAIS or the Stanford-Binet. I suspect it comes down to the realist's statement as to which levels he *cares most about* when he is seeking to find out the truth about the way the world is. I do not suggest there is anything wrong with that. I merely say that if that is *all* it comes to, we should not get terribly concerned about whether a proposed weighted composite agrees less closely with my importance value than it does with yours or with the weighting produced by canonical correlation.

The entire discussion above has been put in terms of the scientific realist. What special concessions must we make to the fictionist? I think none. For the fictionist, all that is "real" is the theory's properties, especially its factual track record, its predictive and technological performance. Therefore we do not expect him to object to the canonical correlations weighting of the track record unless it does too gross a violation to his intuitions *on the instrumental side*. He might complain a bit about our procedure for assigning weights to his set of theory properties on the basis of their correlation with something he considers chimerical (verisimilitude), but given his attitude to that concept, how the realist assigns weights to *it* is, while to a fictionist a fruitless activity, also quite harmless. It is unclear why the instrumentalist would have any reason to assign weights in the first place. If he did so on the basis of a factor analysis of the theory properties, it would be of interest to the realist how the fictionist's "best composite," which now has an internal instrumental meaning alone, agrees with the realist's composite reached by canonical correlation. But I see no reason why either party would be distressed if the composites did not agree. My prediction would be that they would agree quite closely.

You may say, "But, wait, we are talking here about theories *so strongly evidenced* (track record) that they are now enshrined in all the textbooks, not considered to be problematic but settled, usually not even referred to anymore as being theories. And yet you are saying that the fictionist would have *no* interest in an index based upon the relationship between discarded theories, or early stages of good theories, and these 'criterion theories'?" Well, if that is an incoherency, it is not mine. If an incoherency exists there, it flows from the incoherency of the fictionist position, which I do not wish to discuss in this place. Verisimilitude is defined explicitly as an ontological predicate of theories; it presupposes scientific realism and the correspondence theory of truth about the external world. If the fictionist really means what he says, that he does not attach any meaning to such correspondence, or that he does not care about it one way or the other, then he should not bother his head about the weights that the realist assigns

to the *V*-levels. A more detailed and finely honed treatment of the weighting problem is presented in Appendix 1 (pp. 456-462).

PERFORMANCE INDEXES: ILLUSTRATIVE EXAMPLES

From a neo-Popperian or Lakatosian standpoint, the shift from truth to verisimilitude destroys the beautiful simplicity of the Popper *modus tollens*, and accordingly I shift attention from the dichotomous notion of "passing a severe test" to the broader notion of "coming close enough to be a damn strange coincidence," taking a punchy phrase from Wesley Salmon (1984 and personal communication June 1980; Meehl, 1990a). Two crude indexes of empirical performance are presented here without detailed defense or reasons for choice of the mathematical form, as the arguments have been developed at length elsewhere (Meehl, 1990a). Then I will sketch possibilities of two others, giving us a set of four indexes of a theory's adequacy. These four, while only a fraction of the lists of desirable theoretical properties that have been proposed by various authors will suffice to exemplify my cliometric proposal. (For other criteria for evaluating goodness of theories see, e.g., Cohen & Nagel, 1934, pp. 207-215; Copi, 1961, pp. 426-433; Dauer, 1989; Faust & Meehl, 1992; Feigl, 1929, pp. 131-137; 1950, pp. 38-41 [1981, pp. 196-200]; Frank, 1954; Hempel, 1966, pp. 33-46; Kordig, 1971a, 1971b, 1978; Kuhn, 1977, pp. 320ff; Laudan, 1984; Margenau, 1950, pp. 81-121; Newton-Smith, 1981, pp. 226-232; Popper, 1962, pp. 231-233; Schaffner, 1970, pp. 318-330; Shapere, 1977; Watkins, 1984, p. 130ff *et passim*; and see in this article pp. 379-380, 406.) *The merits of the overall cliometric proposal do not hinge on whether the particular indexes I suggest are optimal (surely not!) for their respective purposes or even whether they satisfice*. I will not accept (as dispositive) objections on grounds of imperfection, whether conceptual or numerical. If metatheory is the empirical science of science, we are to appraise a concept, theory, or formula by its performance, how well it "works." If, in the early stages, an index "works" *at all*, the cliometric metatheorist does what the first-level scientific theorist does, he sets about improving it. If it appears not to "work" even a little bit, we try to analyze why and return to the drawing board.

The first index of empirical adequacy relies on the generally accepted notion that when a theory successfully derives a replicable observational fact it is thereby corroborated. [I bypass whether the derivation is before or after the fact, and the well known problems of the extent to which "derived," in the usual sense of the empirical scientist, can be equated with Hempelian "deduce" (cf. Meehl, 1990b).] How strongly this empirical success episode corroborates or confirms the theory depends upon the precision of the prediction. Working scientists for three centuries or more, including our contemporaries who know nothing about philosophy of science and ignore it as

irrelevant, have used language suggestive of Popper's "risky test" or Salmon's "damn strange coincidence," as when they say, "The theory succeeded in making remarkably detailed observational predictions." For reasons argued in the paper just cited, I have concluded that scientific phrases like "remarkably precise," "extremely detailed," or "highly accurate" usually refer to a composite of two quantitative values (judged or measured). The first of these is the narrowness of the predicted interval in relation to what was antecedently possible from background knowledge. "Background knowledge" includes other accepted theory, previous research, and even commonsense or everyday observation. The second is—whatever the allowed interval was in relationship to the antecedently conceivable range—how close did the prediction come? These two numerical values tend to potentiate or offset each other, to such an extent that a theory which successfully predicts a numerical value but which tolerates a large portion of the antecedent range (Spielraum) is not accorded as much "success" as a theory which tolerates a very small interval of the Spielraum but "misses" by a small amount. *This scientific practice accords with a shift of ontological concern from literal truth or falsity to the emphasis on verisimilitude*. Given suitable conventions for defining the Spielraum (S) based upon the background knowledge of a particular scientific domain or subdomain, we may define a corroboration index for a particular experimental or statistical study as

$$C_i = \left(1 - \frac{I}{S}\right)\left(1 - \frac{D}{S}\right)$$

where I = Interval tolerated by the theory, and D = Deviation of an observed value from the edge of the tolerated interval. The product function potentiates "how risky" and "how close" mutually, a Fisherian interaction effect.[20] Considering the worst and best cases, I have suggested a standardization of this index which has the familiar property of lying in the interval (.00–1.00); and the cumulative corroboration index for a theory is then defined as the mean value of the particular corroborations. I leave open whether the corroborations to be counted are those that meet some minimum standard of replication, e.g., two from different labs or clinics. A variant of such numerical point-prediction, almost as severe a test as this, arises when a theory is too weak to predict point-values but is strong enough to predict *some* of the point values from *others* (observed but not theoretically forecast). This is particularly important in the life sciences, where the numerous latent and historical factors causally responsible for the distributions of observable variables preclude even a correct "structural" model (e.g., dominant schizogene) from directly entailing an observable value x_i but from x_i we can infer an

[20]For details, illustrative examples, and criticisms, see Meehl (1990a).

observable x_j. [Cf. my emphasis on *consistency tests* in research (Meehl, 1973, pp. 200–224; 1990a; 1992a; Meehl & Golden, 1982).]

A common situation found in both the inorganic and life sciences is one in which a theory, while too weak to generate numerical point or small interval predictions, is at least strong enough to predict the form of the function relating two quantitative observables.[21] The antecedent Spielraum is harder to specify in this case, but what I suggest (see Meehl, 1990a) is an empirical survey of the function forms that appear in a domain of scientific literature (e.g., mammalian learning, human vision) and assigning as the theory's intolerance the empirical relative frequency with which the function form $F(x)$ that it predicts has appeared in the domain literature. This suggestion may bother mathematicians (and even statisticians), but it seems in harmony with the general approach that metatheory starts with the empirical facts of scientific research. Obviously some such empirical basis must be used, inasmuch as the pure mathematician cannot give us help—the cardinal number of single valued functions being the second uncountable infinity!

One property of a theory commonly mentioned by metatheorists, and in my anecdotal experience universal among working scientists, although not every metatheorist is willing to give it high weight, is the *qualitative diversity* of empirical contexts for which the theory makes predictions. My impression is that in psychology this indicator of performance produces acute discomfort (cognitive dissonance) when it collides with other strongly valued indicators such as the previous two. Thus, for example, psychologists appropriately impressed with the nearly perfect replicability and detailed prediction of curve forms and slopes achieved by the operant behaviorists (cf. Ferster & Skinner, 1957) are troubled when they find these theorists refusing to explain the large body of data collected pre-Skinner by the studies of rats in mazes, especially the research on latent learning motivated by Tolman's expectancy theory (see MacCorquodale & Meehl, 1954). When one turns to the field of language, most psycholinguists believe the statistical data on the development of verbal behavior in children cannot be reconciled with Skinner's theory or, for that matter, any other "associationist" theory of language. I have not yet concocted an index of qualitative diversity, but for the purposes of this paper I will sketch out how one might go about such a thing. Consider the broad behavioral domain within psychology denoted by the phrase "mammalian learning." The different experimental designs that study mammalian learning, or even more narrowly, the learning of the white rat, run into the thousands. Nevertheless it is not an impossible task

[21]A strong theory may also predict the parameters; a weaker theory may not be able to predict the parameters but can say that a parameter in the function measured in one experimental context will appear with approximately the same numerical value in another. Or it may be possible to predict a relation between parameters in a given context (for example, the asymptote and growth constant of a learning curve) despite inability to predict the parameters or to transfer them across contexts.

to compile a rough set of experimental dimensions which define regions in the experimental space, within which regions particular experiments may differ as to the empirical parameters. We can start with a list of properties or dimensions thus: Species? response (locomotion, manipulation, gesture, verbalization)? reinforcement (positive, negative; the appetitive including hunger, thirst, sex, maternal, exploration, "novelty"; the aversive including electric shock, loud noise, etc.; if aversive control, escape versus avoidance, primary or conditioned stimulus)? controlling stimulus modalities? and so on. One is at first dismayed by the welter of possibilities. But I remind myself that, while there are some 18,000 trait names found in the Allport-Odbert (1936) list (culled from Webster's *New International Dictionary*, 1925 edition), such an initially discouraging fact has not prevented personality theorists and psychopathologists from identifying the important major ways in which persons differ in their traits, from developing fairly valid psychometric and rating scale measures of them, from investigating their dependence upon childrearing practices and—especially in recent years—their biological heritability. The coarseness of grouping in the experimental hyperspace for defining empirical regions is itself something to be investigated empirically, and in the first phase of cliometric work in a domain, one might employ very loose categories. Given a categorization or dimensionalization of the experimental space that satisfices, one could make a literal count of the proportion of experimental contexts for which the theory makes derivations (correct or not). A refinement might consist of some system of weighting domains as to the quantitative range or qualitative diversity the domain itself exhibits. Another refinement of such an index might consist of a measure of the "density" or "concentration" and the standard deviation of densities over domains. For example, given two theories each of which has something to say predictively about a dozen domains of mammalian learning, one might be more impressed with a theory whose predictions are fairly evenly dispersed over contexts than with one whose predictions pile up in a single context and which manages to say something about only one or two experiments in each of the other broad contexts.

A fourth commonly invoked criterion of theoretical success is reducibility. One need not be a strong reductionist or "in a hurry" to make reductions to agree that scientists are usually impressed when they find that the concepts and laws of a theory can be reduced to those of a theory at another level of Comte's famous pyramid of the sciences (see Comte, 1830-42/1974, 1830-54/1983; Oldroyd, 1986, Chapter 5; Meehl, 1990b) or that the theory has successes as a reducer of theories standing higher than itself in the pyramid. Reducibility of concepts is a necessary but not sufficient condition for reducibility of laws, although a dogmatic metaphysical reductionist will hold that the only time the former is not sufficient for the latter is where we have incomplete information about the compositional and structural arrangements. Here again a

convention would be set up for distinguishing empirical subdomains; and there seems to be an asymmetry between reducibility downward and reducibility upward, the latter carrying somewhat less weight. There also is a distinction between reducibility in the strong sense of derivability (again, having provided the structural arrangements) and a kind of negative reducibility that involves constraining principles (cf. Meehl, 1990b). Perhaps we do not claim to be able to reduce the concepts or laws at a given level in the pyramid to those below, but at least we are not dealing with a theory which is *prima facie* incompatible with well corroborated theories at a lower level. For example, it is misleading to say that all of the phenomena of meiosis that are involved in the geneticist's understanding of linkage are strictly derivable from the laws of chemistry and physics, which they are not. But we would not countenance a theory of meiosis that says the meiotic spindle consists of tiny silver wires. While there are examples of organo-metallic compounds (e.g., hemoglobin involved in living systems, we are confident that there are no living systems that have, as part of their machinery, silver wires.

Suppose preliminary historical research, based on sampling the literature of scientific journals and textbooks, permits us to conjecture that these four indexes of theoretical performance each possesses some degree of validity, the latter term meaning that they are substantially correlated with the scientific community's long term appraisal of a theory's verisimilitude. We might then make a random or representative sampling (I don't care which, the latter being useful mainly because it reduces the standard error, it does not introduce a new qualitative principle) of "mini-theories" referring to the facts of various scientific subdomains. We might, for example, choose a score of mini-theories from some field such as organic chemistry where a certain mini-problem (say, the molecular structure of benzene, or how photosynthesis works) has had proposed two mini-theories at approximately the same time, and then follow the course of the quantitative change in these four performance indexes over the life history of the mini-domain. To be on the safe side in the early stages of such cliometric work, one would choose mini-theories that have been universally accepted and for the last fifty or more years presented as "facts" (rather than theories) in standard textbooks. One might define an "experimental half-life" for a dead mini-theory by first counting the experiments performed with an eye to appraising it (with or without its competitor) until such time as the theory is abandoned by the community of scientists and is not mentioned in textbooks except as a passing historical comment (as we now mention caloric and phlogiston). If there had been 100 experiments performed on the theory before it was definitively abandoned by almost everybody in favor of its competitor, we could compute each performance index as it stood by the time 50 of them had been reported in the literature. The first thing a psychologist would think of would be to assess each index as to its predictive validity

for the theory's fate in the long run. Dividing our theories into winners and losers, we have a dichotomous "criterion variable" to which we can apply the linear discriminant function (Harris, 1975) predicting this dichotomy. The weight ("importance") assigned to each of the four indicators in this linear discriminant function depends jointly (configurally) on (a) its own predictive validity with respect to the theory's fate and (b) the pattern of its statistical relations with the other three. The resulting set of four statistical weights quantifies each performance index's importance in forecasting the theory's terminal state. All this is discriminative-validating use of statistics, provided "50-year status in textbooks" is accepted as an operational proxy for "final social fate."

It is natural for the psychologist to consider next the relation among these indexes from a more problematic standpoint, involving the causal-theoretical use of psychometrics as contrasted with the discriminative-validating one. Conjecturing that a theory's verisimilitude is (fallibly) reflected in each of these four performance indexes, we may ask whether the pattern of relationships among the four, taken over a large set of theories, is consistent with this conjecture. We compute the correlation coefficient (Pearson r) between the index pairs and examine the six coefficients by Spearman's tetrad difference criterion,

$$r_{12}\, r_{34} - r_{13}\, r_{24} = 0$$
$$r_{12}\, r_{34} - r_{14}\, r_{23} = 0$$
$$r_{13}\, r_{24} - r_{14}\, r_{23} = 0$$

(the special case of Thurstone's multiple factor analysis corresponding to the mathematical fact that the rank of a correlation matrix = 1 if all of the second order minors vanish). Approximate satisfaction (I would not do a significance test, but some would) of the Spearman criterion does not tell us that there is only one single factor underlying the correlations, but it does tell us that one factor *would suffice* to explain them. We also have from this analysis the factor loadings of the four performance indexes, which, speaking causally, corresponds to the relative *influence* of the inferred single underlying factor upon the indexes. We examine the profile of factor loadings yielded by this internal analysis (without a "truth criterion" for the theories) to see whether it is similar to the profile of the discriminant function coefficients where the ultimate social fate of the theory was available as a dichotomous criterion. High similarity between these profiles of factor loadings and discriminant weights would tend to corroborate the notion that the indexes are (fallible) measures of verisimilitude. It is also possible, knowing the factor weights, to construct a composite estimate of the inferred factor; as in the discriminant function we can sum the squares of the discriminant weights to obtain a composite validity coefficient.

I've used these four "performance" indexes to illustrate a general point:

If case studies, whether treated impressionistically in the usual way or tallied statistically, indicate that the community of scientists does in fact appraise theories by a particular theoretical property or relation, then a collection of such properties and relations can be statistically studied while paying attention to their predictive validity with respect to the theory's long-term fate in the scientific consensus and also to the internal relationships of the indicators. Reasonably good agreement between the factor weights (internal, bootstrapsed) of such indexes and the predictive (external, "criterion-oriented") weights would tend to warrant our moving from a purely discriminative-validating view of these weights to a causal-theoretical one, namely, that the theory's performance testifies (in a stochastic sense) to its verisimilitude.

Let me emphasize that the Faust-Meehl Thesis is not tied to any particular statistical method or metatheoretical concept, certainly not to my suggestions here. For example, the numerifying of verisimilitude does not depend on my "Theory Specification Levels"; they are offered mainly as illustrative of an approach different from the logician's. That list "fits" some kinds of theories (input-output theories that are at least somewhat mathematicized) and would be a poor explication of verisimilitude for other kinds of theories. *All* choices of categories, dimensions, and formalism in science are judged by their "fruitfulness," which mainly consists in the resulting orderliness of the empirical relations. A common mistake scholars make when first confronted with our thesis is supposing that they can decide (adversely) from the armchair without having any statistical facts. On the current conception of metatheory as the empirical theory of scientific theorizing, that, of course, cannot be done. Theoretical (or intuitive, commonsensical) objections to a proposed quantitative index are properly viewed as *matters to be concerned about, problems to solve, possible deficiencies to watch for, modifications of an index to consider*—but *not* as dispositive objections, grounds for abandoning the whole idea. As all theories arise in a sea of anomalies and would be "strangled at birth" (Feyerabend, 1965, pp. 24 fn, 229-230, 122 fn, 249-250; 1970, pp. 36-45 *et passim*; 1971) if one applied "instant (*modus tollens*) rationality" (Lakatos, 1970), so any proposed set of metatheoretical properties, dimensions, indexes, composites, and statistical methods of correlating them will have apparent defects. In an empirical science, the way one ascertains whether a concept or formula has any merit is to try it out on data, whereby one *finds out* what mathematical orderliness and conceptual illumination it can provide.[22]

[22]Nickles (1986), forcefully and insightfully criticizing the research program of the Virginia Polytechnic Institute group (Laudan, *et al*., 1986), raises some important questions about how metatheory, when historically oriented, should proceed. He emphasizes, as I do here, the dangers inherent in the case study method (as commonly employed). But I do not discern any signs that he views his own critical theses as empirically testable conjectures or that he thinks the difficulties inherent in the case study method may be partly alleviated by an actuarial approach.

In preceding text I have treated verisimilitude dichotomously, assigning value True or False (1 or 0 in calculating the discriminant function), then inquiring how the profile pattern of discriminant weights on aspects of theory performance resembles that of factor loadings inferred from *internal* relations of the performance measures. We would relate the performance indexes at an early stage (or, say, a theory's "half-life") to its terminal fate—oxygen is in, phlogiston is out; nucleotides are in, proteins are out. Since the entire business is stochastic—*intrinsically* so, not just due to our ignorance or lack of cleverness—that a few of the theories scored 1 in discriminant analysis may some day, to everyone's astonishment, be refuted, does not vitiate the procedure. Here again, philosophers can profitably learn from psychologists. In building and "validating" a mental test (intelligence, personality, vocational interest) we routinely employ "criteria" that we *know* are fallible, if for no other reason than because their measured reliability coefficient is < 1 (and the square root of reliability is an upper bound on validity). Nevertheless, over a period of years we often learn to trust the test more than the initial crude criterion. There is nothing strange about this psychometric bootstrapsing either causally or mathematically.

We can correlate multiple indexes of theory performance at an early stage of research and at half-life with our verisimilitude index, where the "final" theory, enshrined in textbooks, is the criterion. A variety of analyses of the numerous time series immediately suggest themselves.

At a more advanced stage of this cliometric process, refinements of the indexes, both as to nonlinear transformations of each and—probably a more important source of nonlinearity—the appearance of significant configural effects (Meehl, 1954/1996, pp. 131-134) would be investigated. As intensive case studies accumulate, selections of particular episodes in the history of the science would be partly motivated by our identification of those aberrant episodes in which a high validity (but still fallible) composite fails to forecast the theory's long-term fate. We recognize that there *should* be a small subset of such anomalous episodes that will not be revealing in any systematic sense, because they will be literally attributable to "chance," in the same way that the psychometrician knows that forecasting events or course of individual lives (e.g., criminal recidivism, success in a profession, recovery from psychosis, longevity, cause of death, number of children) may be a matter of "chance" (Meehl, 1978, p. 811). Prediction of an external criterion aside, the classical psychometric formalism regularly decomposes the variance (e.g., of a mental test) into shared components (general factor, group factors), specifics (to each test alone), and error (unreliability, undesignated "chance" factors). There will be an ineradicable stochastic element in a theory's track record, no matter how successful the metatheorist may be in formulating guidelines, and the working scientist will be flexible in following them. Sampling from the great universe of facts, there is

of course nothing that the scientist can do, *with or without the metatheorist's helpful advice*, to guarantee that the most decisive facts from the domain of experimental contexts will be sampled.

Nevertheless, we would expect to discern, with the aid of intensive case studies brought to bear on the unsuccessful predictions, special features of these episodes that, had they been attended to, might have led to a different assessment or at least alerted the appraising scientist to heightened danger of a misprediction. Here again, the combination of the statistical approach with the intensive case study may be expected to suggest amendments in the conceptual (verbal) formulation of a performance criterion and hence to corresponding modifications of the index used to quantify it. In the sort of statisticized metatheory I envisage, paying attention to metatheory (i.e., listening to the helpful advice and tentative guidelines offered by the historian and philosopher of science) would no longer depend (as it does today) on the biases, tastes, and interests of the working scientist. The theorist and theoretically motivated experimenter would ignore metatheory at his peril, just as today a biologist would ignore, say, Fisher's statistics at his peril. A scientist dealing with a subject matter in which probability plays a major role, as in genetics or psychopathology, cannot today ignore the advice of a statistician on the ground that "I am a geneticist, and I'm just not interested in mathematical statistics." It will be a historical meta-test of the maturity of metatheory that working scientists will, by and large, act in conformity to its guidelines, always remembering that there will be a small number of anomalous successes achieved by those methodological nonconformists who are either genius mavericks or just plain lucky.

DESCRIPTIVE VERSUS PRESCRIPTIVE GENERALIZATIONS

As stated in my opening paragraph, to find a rigorous philosopher of science saying that his discipline is a branch of social science (Sneed, 1976) comes as a shock to one raised, like myself, in the tradition of logical positivism. I can offer no statistics from a Gallup Poll, but I dare say that even among the younger generation of philosophers, only a minority (how small?) would subscribe to that classification of their enterprise. There is a trivial sense in which the statement is correct, namely, science is concocted by humans, is a product of the workings of mind in society, and therefore when the metatheorist studies scientific theorizing he is "obviously" doing social science. Putting it that way, I am fairly comfortable with it. But it is not that obvious categorization that I think most philosophers find objectionable. What they dislike and fear, as I do, is that this harmless looking truism is sometimes taken to mean that the traditional job of the philosopher, that of philosophical *analysis*, with an eye to the aims of rational reconstruction and justification, should be liquidated in favor of purely empirical psychosocial description. I don't know how many of the younger generation of metatheorists hold that position,

or whether all of those who do are logically consistent in the way they do it; but some do talk and write in that way, so I want here to explain why I consider it a grave mistake.

Let me begin with Reichenbach's well known (and, today, sometimes repudiated) distinction between the *context of discovery* and the *context of justification* (1938). I take the old-fashioned position that is it imperative to begin by making that distinction, and to understand why one makes it, after which we can proceed to qualify it and fuzz it up a little. A total failure to make it would allow as legitimate what beginning logic courses regularly label material fallacies, such as the *argumentum ad hominem*, the genetic fallacy, the disparagement of an opponent's motives in an argument, and the like. Anyone is simply confused who does not see the difference between (a) giving a psychological or sociological account of how Jones came to concoct an idea before Smith did and (b) inquiring whether Jones's evidence and arguments are better or worse than Smith's. It is interesting for the historian of science to know the story of Kekulé's dream of the hoop snake; but if an organic chemist argued against the theory of the benzene ring structure because "it was based upon a dream," or someone argued in favor of its truth because it was a fine example of creative "intuition by analogy," that person would be read out of the chemists' scientific club. [An excellent discussion, analyzing and defending Reichenbach's distinction, is given by Siegel (1980).]

Part of the problem with Reichenbach's two contexts is that people wrongly infer, from this fundamental distinction between two kinds of discourse having different goals, that no statement that can occur appropriately in one of the two contexts should appear also in the other. Reichenbach did not assert this and, so far as I can make out on re-reading him, nothing that he said about the two contexts implies it. The two contexts are defined by the *aim* of the discourse: what question is being asked and what kind of answer would be acceptable. Nothing about that threshold distinction entails that if a sentence appears appropriately in the context of discovery, it could only appear inappropriately in the context of justification, or vice versa. *Example*: "What psycho-social influences led MacCorquodale and Meehl to attempt a formalization of Tolman's expectancy theory of animal learning?" (MacCorquodale & Meehl, 1953, 1954). This is clearly a question in the context of discovery, and it invites a psychological and sociological sort of answer. However, that psychological and social answer has a *content* that includes certain logical and methodological relations, thus: "They were both Minnesota PhDs and educated in a strongly behavioristic, positivistic, and psychometric orientation, from which vantage point Tolman's avowedly fuzzy concepts required cleaning up to be scientifically acceptable and to derive, in any rigorous way, experimental predictions." It is obvious that if one merely said, "Well, they both

received their doctorates at Minnesota," that would not be an adequate explanation except to a listener who knew how to fill in the above cognitive content. *Example*: "Why do you continue holding Fisbee's theory when Hocheimer's experiment clearly refutes it?" This is a question about reasoning and evidence, it sets up the context as one of justification, it requires a rational answer within the principles and rules of the scientific game. And yet, the correct answer sounds as if we were in the other context, if one mistakenly supposes that the legitimate statements in the two contexts must be disjoint, the valid and adequate rational reply in this instance being, "I know that Hocheimer has a terrible bias against Fisbee's theory, he has a father surrogate complex about Fisbee, who was a tyrannical PhD advisor, and as a result his experiments are biased in subtle ways. So I do not receive his protocols into the corpus." I am confident that Reichenbach would have been perfectly contented with both of these answers.

The easiest and soundest way to approach this problem of the factual and the normative, the descriptive and the prescriptive, the "is" and the "ought," the empirical generalization and the rule, the actual and the ideal, is to say that in metatheory we attempt to formulate hypothetical imperatives about the scientific aim to get credentialed knowledge. "If you want to get knowledge of the scientific sort, that possesses the kind of credentials that science often provides and always aims at, you should, by and large, do so-and-so; and you should, by and large, avoid doing such-and-such. This advice is based upon empirical facts from the history of the development of scientific knowledge." Not being an intellectual fascist or mind controller, I have no urge to coerce a Hindu mystic or an Asbury Heights anti-scientific hippy to play the scientific game. But if someone claims to be playing it, I may see fit to offer some advice, either as scientist or part-time philosopher.

It may be objected, if what we have is simply a hypothetical imperative relying upon certain historical generalizations concerning "success" (I here bypass the problem of circularity in that kind of argument), have we not thereby eliminated all of what has constituted traditional philosophizing, the analysis of the structure of knowledge that is *explanatory* and *justificatory*? I say not. I say that viewing metatheory as the empirical theory of scientific theorizing, if it is to be "pretty much like other empirical theories" (as its highly empirical advocates claim), will have as one of its aims to *explain*, to *understand*, to *rationally reconstruct*, why science succeeds; and, equally important, to explain why science fails when it does. To make this plausible I shall consider some simpler examples not involving metatheoretical reconstruction.

I start with an animal that, while it seems to have cognitions in a suitably defined sense, does not concoct abstract theories and certainly does not engage in metatheory

about its own cognitions, namely, the white rat in the Skinner box. In his classic work *The Behavior of Organisms* (1938) B.F. Skinner showed that rats can develop stable behavior if rewarded with a food pellet after making 192 lever pressings, the *fixed ratio schedule* denoted FR192. But to reach that level Skinner had to proceed by degrees, increasing the ratio gradually. If a rat on a continuous reinforcement schedule were shifted suddenly to FR192, its lever pressing behavior would extinguish; and in unpublished work not mentioned in the book for obvious reasons, Skinner (or W. T. Heron?) showed that a rat could be starved to death in the box even though, had it responded under the lean schedule FR192, it would have been calorically ahead of the game. Now if someone asks about the death of such a rat why it died, a satisfactory explanation of death (despite the availability of food if it had emitted the right behaviors) will unavoidably refer to various mathematical relations between behavior dispositions and the "rule" imposed by the experimenter in programming the Skinner box. These include some rather complicated proofs about statistical probability. For example, the reason the cumulative record has the shape it does on FR192 lies in the initial nonrandom clumping or grouping of responses, which gives rise to a tendency for a series of lever pressings closer in time to acquire the character of a discriminative stimulus S^D as contrasted with a low rate of responding, which becomes an S^Δ. Nobody would object to this analysis on the grounds that Skinner had claimed to be a behaviorist offering a purely empirical theory; and yet here he was employing theorems from the probability calculus, which were used to relate in a somewhat complicated manner the propensities of the white rat on the one side with the "causal texture of the environment" (as Tolman called it) on the other side; and he included counterfactual statements such as, "If the rat *had* continued to press when first shifted to the high ratio, it would have been able to sustain life."

Consider a second example: A strongly "empirical" economist (they seem to differ very widely in this respect) undertakes an investigation of business firm bankruptcy. He obtains a random sample of business enterprises that were successful and of others that went broke, and he calculates statistics on attributes of the chief executive officer (CEO), Chairman of the Board, Plant Manager, etc. He summarizes his findings by presenting actuarial generalizations of psychosocial factors that appear to be adverse to a firm's success, e.g., the CEO is a problem drinker, the Chairman of the Board is going through a complicated and painful divorce proceeding, the average IQ of the board members is lower than that of board members at competitor firms also manufacturing widgets, the Plant Manager has a peptic ulcer, and the like. Suppose he offers the statistics as descriptive and explanatory, as enabling us to understand why some firms tend to go

bankrupt and others do not. Would we feel intellectually satisfied with such an explanation as to *why*? I certainly would not, and I do not think most economists would. And why not? Simply because some intervening stages of the expected *explanation* have been left out, and they are crucial for any genuine intellectual comprehension of the causal sequence. Someone naive (or pretending to be for purposes of criticism) could ask the economist, "What has an executive's divorce got to do with the sale of widgets? Presumably 99.99% of consumers don't even know his name, let alone that he's having an unpleasant divorce. You have not given me a clue as to why this irrelevant fact tends to produce bankruptcy." And so with the other three psychosocial facts I have listed as showing up in the statistical summary. The lacuna in the proffered "explanation" is, of course, that these psychosocial facts tend to produce irrational decision-making and *that* is what is directly responsible for the bankruptcy. If we had other research showing that suffering from peptic ulcer, while very unpleasant and perhaps life threatening, is not in fact correlated with the quality of executive decisions, we would tell our economist that his statistics must have something wrong with them, or are artifactually reflecting some third variable, that this particular item—although apparently in his study a *correlate* of bankruptcy—cannot function as an *explainer* of it. If the CEO is an alcoholic, the reason we will accept that as an explanation is that we know from other evidence and common life experience that one's judgment and memory are likely to be impaired after a five martini lunch. I need not belabor the point, which is simply that, in order to fill in the causal chain so that this list of psychosocial correlates of bankruptcy can function as answers to the question "Why?", one must get past ulcers and divorces and booze and IQ to the next links in the causal chain, *and the descriptions of those links must involve distinctively economic concepts and references to decisional rationality*. If you do not talk about accurate estimation of the competitors' advertising, economies of scale, marginal costs, probability × utility, "rational expectations," consumer preferences affected by a recession, and such distinctively economic conceptions, you do not have an explanation at all.

I submit that the empirical metatheorist is in precisely the same boat as the animal psychologist and the economist, although his domain is scientists inventing, appraising, and modifying theories, rather than white rats pressing levers or business executives making decisions. To the extent that the working scientist, whether philosophically oriented or not, tends to follow rules and principles, and in argument tends to invoke them, even a "purely descriptive" account of scientific behavior will be gravely defective if mention of rules and principles of rationality is prohibited from the discourse. The scientist being a rule-obeying and rule-guided and often rule-mentioning animal, to forbid the metatheorist to inquire into the structure and connections of a set of rules and principles, including their appropriateness given the stated aims of the rule user,

seems a strange piece of advice. The point is not merely that we have no compelling reason to forbid metatheorists to talk about these matters; rather I assert the stronger thesis that the metatheorist cannot carry out his defined job, not even *describing* scientific thought let alone *explaining its success*, if he excludes considerations of rationality, logic, formal and material fallacies, mathematics, and probability theory. If we are clear about this, the philosopher interested in traditional problems of analysis, justification, and prescription will not feel nervous about the new definition of metatheory as a kind of social science.

While there is no reason to exclude use of the probability calculus from metatheoretical reasoning any more than from sciences such as physics, genetics, or epidemiology, some may be unhappy about the distinction between the *epistemic* meaning of probability and the *physical frequency* meaning of it. Whether "in principle" all probability statements can be translated without residue into statements about the limit of relative frequencies remains in dispute. What does seem tolerably clear and not dependent upon the resolution of that issue, is that humans—both scientists and other people—constantly rely on "probability as the guide of life" in situations in which no statistical frequency is available to them and even in situations in which it is difficult to formulate a clear conception of what such a relative frequency would consist of. Even if one accepts unreservedly Carnap's (1945) distinction between probability$_1$ (degree of confirmation or evidentiary support of a hypothesis) and probability$_2$ (the limit of a relative frequency of events or properties in some defined physical collective), I believe nobody has maintained that these two, however conceptually distinguishable, are uncorrelated. *Example*: Suppose some genius super-Carnap of the future were to succeed in constructing a general algorithm of inductive inference, one that was applicable to natural and scientific languages rather than only to the ideal language of Carnap's state descriptions. (I myself find it difficult to conceive such a thing but for the moment I will imagine it to be possible.) Then even bodies of evidence of a nonfrequency sort that we today have to "process" and "integrate" to reach a subjective judgment of confidence could be treated in a formalized, mechanical manner so as to yield a numerical value for probability$_1$. In teaching I have found the appraisal of evidence in courts of law a useful pedagogical device to convince statistically minded students that there is such a process as *rational appraisal of evidentiary support*, such as Carnap had in mind in introducing his probability$_1$ concept, which does not proceed by tallying various properties of events in a reference class, computing their relative frequencies [= p_i values], and operating on such p_i values with the rules of the probability calculus like gamblers, epidemiologists, geneticists, or insurance actuaries. Despite the lack of such *initial numbers* and *combinatorial algorithms*, we do consider the (judgmental, subjective, informal)

appraisal process as *rational*, or as criticizable when it is not. The following list of evidentiary findings that Bruno Richard Hauptmann was the kidnapper of the Lindbergh baby is taken from an elementary logic text:

1. The kidnaper's ransom notes indicate their author was a German, as is Hauptmann.
2. The ladder used to reach the baby's nursery was made by a man accustomed to fashioning wood joints expertly. Hauptmann is a carpenter.
3. The lumber used to make the ladder was traced to the National Mill Work & Lumber Co., in the Bronx. Hauptmann worked there, bought lumber there for neighborhood jobs.
4. The nails used in the ladder are said to have the same grooving as nails of the same size found in Hauptmann's home.
5. The print of a shoeless or wrapped foot outside the Lindbergh home is "similar" to Hauptmann's footprint.
6. The writing on the ransom notes has been identified by an expert as Hauptmann's.
7. Paper like that used for the ransom notes has been found in his home.
8. Hauptmann worked near the Lindbergh home in Hunterdon County, N. J., not long before the kidnaping.
9. An automobile seen near the Lindbergh home shortly before the abduction was the same make, model, and color as Hauptmann's.
10. The kidnaper apparently injured a leg in making his getaway. Hauptmann walked with a cane a few weeks after the crime. About 10 months later he was treated by a doctor for chronic inflammation of the legs.
11. The kidnaper and the recipient of the ransom were one and the same because the writing and signature on the ransom notes and that left in the nursery are the same, and because the extortioner delivered the baby's sleeping garment to prove his "right" to the ransom.

There are these additional reasons for believing Hauptmann got the $50,000:

12. A gasoline station attendant identified Hauptmann as the man who gave him one of the ransom bills, leading to his arrest.
13. Hidden in the garage by Hauptmann's home was $13,750 of the ransom loot. In his pocket was $20 more.
14. A taxi driver identified Hauptmann as the man who gave him $1 to deliver a note to "Jafsie," Dr. J. F. Condon.
15. Dr. Condon dealt with a man who had a German accent, as has Hauptmann.
16. Hauptmann quit his job the month the ransom was paid, opened a brokerage account and spent money on hunting trips.
17. By his own word he lent $2000 to Isador Fisch, the man who he contends gave him the ransom money to keep.
18. His wife quit her job in a bakery and made a trip to Germany.
19. His reluctance to answer questions is viewed as "consciousness of guilt."
20. Hauptmann has a criminal record dating back to his days in Germany.
21. The footprint left by "John," who got the ransom, closely resembles that of Hauptmann.
22. On a board in his home officers found Dr. Condon's address and telephone number penciled. Other numbers, including that of a ransom bill were there.

23. Hauptmann said he got the telephone number from a newspaper. It was not published.
24. Hauptmann said the bill number was that of one given him by Fisch, but it was before the time he said he first met Fisch [Castell, 1935, pp. 207-208].

Suppose a metatheorist interested in legal inferences and in the relation between the two concepts of probability were to form a collection of a thousand murder cases in some of which the defendants were acquitted and in others found guilty. And suppose these cases were chosen on the basis of the fact that an almost gold standard criterion subsequently became available. Thus, for instance, assume the defendant, or somebody else not put on trial for a certain murder, made a death bed confession that he had done the deed, and told the priest (presumably he would not have the priest if he did not believe in the religion) where he had buried a tin box containing such-and-such government bonds along with the weapon that had been used in the killing. After his death we dig up the box at the place described and there are the bonds, and there is the weapon with his fingerprints on it, ballistic markings matching the fatal bullet, and so on. It doesn't matter whether a minuscule percent of even these gold standard cases are somehow, by some inconceivable mischance, erroneous. Surely we are .99 confident under this kind of evidence who did it. Thus we know in which instances the jury found properly and in which instances they erred. We apply the super-Carnap genius's inductive algorithm to the evidence introduced, such as the Hauptmann list of facts above. This generates for each case an "objective" probability$_1$ number quantifying the degree of epistemic support. We could then arrange our thousand murder trials in order of this confirmation index. Suppose now it turns out that there is no correlation between the value of the confirmation index and the probability of a correct finding. What would we say? We would say that the super-Carnap was a very clever fellow and had made a nice try but had failed. If a "properly assessed" probability$_1$ has no tendency to provide fair betting odds, which it cannot be doing if it fails to correlate with long run frequencies, it is radically defective, no matter how plausible the *a priori* reasoning behind it. In this deep sense frequentists such as Reichenbach have a valid point about the privileged status of their concept of probability. Even if the "logical," "semantic," "epistemological," "evidentiary" kind of probability$_1$ is *conceptually* distinct and rarely *computable* from empirical frequencies, it is nevertheless linked to probability$_2$ in this criterial fashion. If a nonfrequency method of appraising evidentiary support leads me to erroneous conclusions in the long run or enables me to do no better than flipping pennies or entering a random number table, it stands condemned by its track record.

This descriptive/prescriptive question and the section *infra* on "Psychologist and logician" both touch on a core issue in controversy about the so-called "strong program" of the Edinburgh School (Barnes, 1974, 1977; Bloor, 1976), and readers acquainted with sociology of knowledge theory may wonder why I do not discuss it given my acceptance

of the view that metatheory is a kind—although a *special* kind—of social science. My reasons are that (a) I know too little of it; (b) I am not sure I understand it; (c) what I do understand I consider epistemologically incoherent; and (d) I know of no researcher in the developed sciences who subscribes to it, or takes it seriously, or even finds it interesting. For a succinct, fair-minded treatment of the position I recommend Newton-Smith (1981, Chapter 10, "Strong Programmes"). It should be noted, however, that in his otherwise excellent analysis the only causal source of irrational beliefs he *mentions* is "the distorting effect on [one's] judgment of specific interests" (p. 258); whereas I accept the thesis of Nisbett and Ross (1980) and others that cognitive nonoptimality is a basic property of our minds, not always produced by motives and affects of personal psychopathology or biased by group identification or advantage (class, race, gender, nation, party, religion, school, occupation, etc.). One aspect of the strong program controversy that is hard for a psychologist to get excited about (or even to comprehend) is the question whether causal accounts of unsuccessful science, irrational scientific developments, deviations from the prescribed scientific method, are of a *qualitatively different sort* than accounts of "rational" or "truth-reaching," hence successful science. To a psychologist (if consistent), *logical reasonings* are psychisms, they occur in human persons and their rules are socially learned. The collecting of evidence and the marshaling of arguments are biological processes, they are as much a part of "Nature" as our digestion or thermoregulation. (I never heard my logical positivist teachers deny this truism of the naturalist world-view, although they are sometimes accused of it!) The doctrine of some ordinary language philosophers that "reasons cannot be causes" cannot be allowed by a psychologist. The Platonist meta-mathematician may legitimately ask, "In what realm of being is Goldbach's Conjecture true or false, if we never prove or refute it before the sun burns out?", and I am not competent to address that deep question. Fortunately the cognitive psychologist need not do so, since—"wherever it is" that *deducibility* has its being—*deduction* is a process, it takes place in a mind. To the ordinary language objection "Reasons cannot, by their very nature, be causes," the psychologist replies "Whatever mysterious realm (Plato's heaven?) an abstract valid reason may inhabit, the *stating* of a reason, the *tokening* of a proposition, the *uttering* of a persuasive sentence, the *thinking* of a syllogism, are spatiotemporal events in the world and can function as efficient causes." When a school child computes a sum, certain physiological events occur in his sense-organs, brain, and muscles. He has acquired the rules of arithmetic by processes of social reinforcement. Of course, we may take off our psychologist's hat, put on our mathematician's hat, and inquire whether the addition was done *correctly*. He "gets the right answer" because he has learned how to do arithmetic.

If he gets it wrong, we may shift hats again, seeking an explanation which might involve perturbing emotions (e.g., test-anxiety) or motives (e.g., defiance of teacher), but it need not.

Philosophers concerned about the relation of psychological determinism to rationality are tempted to write as if psychosocial causal factors are always the *irrational* ones (e.g., one's Oedipus complex, or race prejudice, or a bout of indigestion from the fried eggs of breakfast). But there is no good reason to so confine the psychosocial causes. We must keep in mind that *rational* thinking, *accurate* perceiving, *correct* numerical calculating, *valid* arguing, are also part of our mental equipment that has arisen by the influence of social learning experiences on our genetic endowment. I may syllogize fallaciously because of a bias or plain carelessness. But if you tell me I have committed an Illicit Distribution of the Major, and I, recognizing your criticism as valid, revise my thinking, *that* (social!) process of communication is only possible because both you and I went to school and took a class in logic.[23]

THINKING CLIOMETRICALLY: EINSTEIN'S PREDICTION OF LIGHT BENDING

When one clearly understands that the inherently stochastic nature of metatheoretical principles requires that they be appraised and weighted actuarially, philosophical argument from case studies is seen in a new light. On almost every page of such books and articles one thinks, "Well, that's a plausible principle, on this example; the 'theoretical' argument seems persuasive; *but what would the cliometric statistics show*?" *Example*: Many (most?) working scientists say they are more impressed when a theory forecasts a novel fact than when it merely explains a fact known (and used) by the theorizer.[24]

---

[23]That strict determinism is incompatible with rationality, as argued by such an eminent philosopher as Sir Karl Popper, I believe I have refuted (Meehl, 1970). Criticism of a related thesis held by the logical positivists, that the only alternative to psychological determinism is "pure chance," can be found in Meehl (1989a).

[24]I did an informal pilot survey of 38 research-productive scientists at the University of Minnesota, mostly of high distinction (including several Regents' Professors and members of the National Academy of Sciences), nearly half of them psychologists, but also scientists in the fields of biology, chemistry, economics, genetics, geography, geology, physics, political science, psychiatry, and sociology. Each was asked:

> "Consider a theory T and 3 facts that are derivable from it, assuming unproblematic auxiliary hypotheses. Two knowledge situations could obtain, given fixed $[T, f_1 f_2 f_3]$ as content:
>
> Case I "Convergence + prediction": $\begin{cases} f_1 f_2 \text{ were known} \\ T \text{ concocted to explain } f_1 f_2 \\ f_3 \text{ ("novel fact") predicted from } T \end{cases}$
>
> Case II "Pure convergence": $\begin{cases} f_1 f_2 f_3 \text{ were known} \\ T \text{ concocted to explain } f_1 f_2 f_3 \end{cases}$
>
> Do you have a preference, as to whether one case gives you greater confidence in the theory? Needn't explain [your answer]."

Philosophers and historians of science differ widely, both as to whether this is usual scientific practice and as to its rationality. Some think it crucial, that a theory is hardly scientific absent such successful prediction. Others (e.g., myself) give it weight but also admit after the fact explanation as evidentiary, although less so than forecasting. Some (e.g., Mayo, 1991) consider it merely a proxy or correlate of severe tests, without separate value absent severity. Carnap thought it irrelevant[25] (cf. Giere, 1969, 1983, 1984, 1988; Howson & Franklin, 1991; Kelly & Glymour, 1989; Maher, 1988, 1990; Mayo, 1991, and references cited therein; Meehl 1990b; Murphy, 1989, and references cited therein; Popper, 1959, 1962; Worrall, 1985, 1989).

The influence of the Strong Actuarial Thesis on one's thinking is exemplified by the most cited historical example, the 1919 eclipse test of Einstein's general relativity theory. This appeals as a nice test case of metatheoretical predictivism because it plays a role in a great Kuhnian revolution, because it (allegedly) "converted" most of the scientific elite in a very short time, and because it led Popper to his falsificationist insight.

Stephen G. Brush, the distinguished historian of science, employs the case history method on this episode to refute predictivism (Brush, 1989). He quotes six physicists as giving heavy weight to the novel, unexpected fact of light bending being *forecast*, but four of these did not repeat this claim in their later writings. Tallying over a score of physicists writing after 1920, Brush finds a strong majority believing they were as much or more strongly influenced by Einstein's derivation of the Mercury perihelion anomaly, which had been known for over half a century. Richard Tolman was the only one who (after 1923) continued to lay stress on the eclipse forecast. This case study clearly refutes the metaprinciple: "All scientists always give greater weight to a theory's forecasting a novel fact than to its explaining a previously known fact." That perhaps deserves refuting, but did any predictionist ever assert it? I believe not. We must of course distinguish between predictivity as a qualitative metatheoretical criterion of admissibility or "scientific-ness" (as in Giere or Popper) and the *quantitative* principle that, given

---

Of the 66% who responded, all but one subscribed to the "predictivist" (Howson & Franklin, 1991) view of evidentiary support. Though this is admittedly a small and geographically localized sample, I have no reason to think the University of Minnesota is atypical in this instance, and I find their near-unanimity reassuring, though one wonders what was going on in the minds of the 34% nonresponders. It is worth noting that, although the behavior sciences have often been singled out for predictivist criticism, all of the psychologists responding ageed that Case I was preferable. One sociology professor also presented the question to 12 graduating seniors in a comparative social structures class. With no prompting, they all preferred the situation involving prediction: "They just felt they would have more confidence in the theory."

[25]I vividly recall his asking, when I was urging Popper's predictivism, "But, Meehl, how can the mere date of verifying a proposition e affect its logical relation $p(h|e)$ to hypothesis h?" Putting the question this way, I had no answer as a matter of formal logic or semantics. Today I would reply, "It can't, which suggests that methodology of science involves more than formal logic and semantics."

admissibility, a forecast novel fact earns more credit than it would if explained *post facto*. One may insist that a scientific theory be capable of deriving novel predictions without holding that succeeding at one such always counts more heavily than all other theory properties, singly *and collectively*. This is like the industrial psychologist's mixed model (conjunctive + regressive) mentioned *supra*: the predictivity property is a successive hurdle for *T*; but once that hurdle is passed, the property now recurs as one among many variables in a long regression equation, susceptible to countervailings by the other appraisal properties. Surely no metatheorist, however strongly insisting on predictivity as a required scientific property or as deserving special weight, has held that no other property of theories should receive any weight at all. If any philosopher were to advance such a preposterous notion, holding that *all other properties* (numerical precision, riskiness, qualitative diversity of fact domains, derivational rigor, independently tested auxiliaries, plausible *ceteris paribus* clause, conceptual "depth," mathematical beauty, coherent fit in Comte's pyramid, minimal parameter adjustment, observational/theoretical predicate ratio, fact/postulate ratio) *are irrelevant*, no one would take him seriously. When predictivity is stated qualitatively, "Prefer a theory that forecasts a novel fact…" it *always* (I cannot think of an exception) requires a ceteris *paribus clause*. In the present instance, *cetera* are not *paria*. For instance, as Brush points out, the light bending numerical fit was not very precise, whereas the perihelion value was known—and derived—with high precision. The light bending derivation did not involve the "deep" components of the theory, as did the perihelion solution. (Eddington admitted that the eclipse result only confirmed Einstein's law of gravitation, not the General Theory.)

The actuarialist (especially the psychologist) would raise a number of skeptical questions about the generalizability of this case study:

How representative is this sample of scientists as to age, prestige, nationality, previous familiarity with the theory, relevant mathematical competence, previous concern about the Mercury anomaly, whether physicist or astronomer, etc.?

In general, when a scientist explicitly states that one argument influenced him more than another, how reliable are these introspections?

If a scientist explicitly states that a consideration weighed heavily with him, and some years later omits to repeat this, how often does this mean he changed his mind? Is that probability correlated with whether the theory's status has meanwhile changed (e.g., from a strange, radical idea to a widely accepted one)?

What proportions of the predictivists and nonpredictivists were aware of Eddington's tendentious reliances on the three sets (two at Sobral, one at Principe) of readings?

What proportions of the two groups (predictivists/nonpredictivists) were aware of (or could have suspected) the conceptual and mathematical weaknesses in Einstein's derivation of the light displacement (Earman & Glymour, 1980)?

How typical is General Relativity as an instance of Kuhnian revolution? (One suspects it was the "biggest ever," given the super-paradigm status accorded to Newton's theory.)

Assuming *arguendo* that Kuhnian revolutions are a genuine taxon, does predictivity have the same weight in revolutions (or crises preceding them) that it does in "normal science"?

What are the relative weights scientists give to predictivity versus numerical precision when they countervail each other? (I conjecture that this is the most important issue presented by the case study.)

Depending on how these questions are answered,[26] the descriptive content of this case study could vary from some such narrow, ungeneralized statement as "On the eve of the biggest revolution ever, the majority of a small nonrandom elite sample of mostly British and German scientists believed themselves to have been more influenced by a new theory's precise derivation of a long-standing clear anomaly than by its less precise and doubtfully valid forecasting of a novel fact," to the strongest "Physical scientists are always more influenced by a theory's precise explanation of an old anomaly than an imprecise prediction of a new one."

Professor Brush does not, as a historian, render a philosophical judgment as to the "soundness" (validity, rationality, fruitfulness) of predictivism, although he does make the interesting and important point that a half-century of failure to explain the Mercury anomaly should properly heighten skepticism as to its Newtonian feasibility. What can the metatheorist, aiming to pass from description to prescription, take from this case study? Not very much, I think. At the "weak" end of the descriptive generalizations continuum, hardly anything; and even the "strong" candidate stated above prescribes or proscribes little. One might think that because general relativity became a conspicuous and (largely) undoubted success story, the instance must prove *something* or other prescriptively, e.g., a "sound policy" of emphasizing precision over forecasting. But it can't even do that, *because both facts were favorable to Einstein*. We do not have here an instance of T_1 deriving (approximately) an "old fact" and in competition with T_2 forecasting (approximately) a novel one. To make matters more interesting, we would want, say, Einstein to *forecast* f_1 *approximately* and a competitor theory to *explain* f_2 *precisely*. As a matter of cognitive psychology, a generalization saying (roughly) "Scientists tend, most of the time, to experience larger increments in credence of T from a more precise explanation of a known fact than from a less precise forecasting of a novel one" is interesting and itself deserves psychosocial explaining by behavior science. As it stands, it is incapable of generating a decision policy, for two reasons. First, the comparative increments of credence in a given theory provided by facts f_1 and f_2, both favorable, is not usually a decision problem. Second, on the rare occasions when it is,[27] *should* a predictivist policy prevail? That sort of principle requires warranty by a

[26]None of them can be answered without statistical analysis of *classes* of *episodes* in history of science.

[27]It is hard to think of any such, but one would be whether to continue work on T if a previously forecasted fact subsequently fails to replicate. A "credence threshold" might be passed here that would not be passed if one had assigned lesser weight for a nonforecasted, now nonreplicated fact.

combination of actuarial data (long term success frequency of the policy) and theoretical rationale "explaining" the policy's statistical success from considerations of cognitive science, logic, probability theory, and armchair epistemology.

The contribution of theoretical meta-meta-arguments is threefold: (1) suggesting what facts to actuarialize, (2) suggesting what statistical analyses to perform, (3) *explaining* success and failure in a rational reconstruction. Whether some rock-bottom desiderata for minimum "rationality" should be allowed to prevail against (apparent?) empirical success, I leave aside. Not being a mystic, I do not expect any such collisions to occur between fundamental rationality (e.g., formal logic) and actuarial success. Of course, *individual episodes* of "irrational" (= unprincipled) success are to be expected. As Lakatos says, even sticking to a degenerating program may sometimes be all right, so long as one does not falsify the track record. Given the stochastic nature of the empirical enterprise, we know in advance that the scientific community's *sampling of the factual domain* ("which experiments to perform," "what to look at") will sometimes result in *misleading evidentiary situations*. In such cases, the scientist who opts for an actuarially suboptimal policy—for whatever reason—may be the "lucky winner."

The "pure case" of predictivism must hold the theory and fact set [T, f_i, f_j, … f_m] constant, asking whether a probative increment occurs when one or more of the derived facts is forecast rather than merely explained and used in concocting T. As descriptive of practice, this question can be researched experimentally, using scientists as subjects, and statistically, using the literature and personal documents.[28] As this pure case does not involve competition between two theories, it may seem of little importance as an empirical bias for adopting a *pre*scriptive principle of theory appraisal. However, we do not know whether intertheory comparisons, as stressed by Lakatos and Popper, present the only important decision problems. When a received theory is firmly entrenched and generally well corroborated by impressive derivational success but is beginning to show cracks in the edifice (e.g., a few numerically strong, replicated anomalies recalcitrant to our ingenuity), there may be numerous individual decisions that never surface in the literature, or even in correspondence, that collectively influence the course of events.

[28]When cliometric metatheory becomes generally accepted and its value appreciated, scientists will become accustomed to the idea that in their capacity of journal referee or project peer reviewer, they may on rare occasions be serving unknowingly as guinea-pigs, whose judging behavior is being studied metatheoretically. Thus a study by Atkinson, Furlong, and Wampold (1982) showed that a journal editor accepted or rejected a paper depending on whether the results were purported to be statistically significant or not. Mahoney (1976, pp. 92-95; 1977) found that different referees tended to accept or reject identical manuscripts depending on whether the results confirmed their own opinions.

"Shall I invest time in trying to patch this theory up?" "Would another replication of Fisbee's experiment make a difference?" "Can this formal derivation be carried through in a different way that does not rely on problematic idealization X?" Even without a live option T_2 to look at, one suspects mini-decisions like these, made quietly by scores or hundreds of scientists, are influenced by the faith they put in T_1's predictive track record. If predictivity is receiving more weight than it deserves, working scientists—not just metatheorists—need to find that out. That a theory is never abandoned until a better one becomes available is itself a generalization needing test. That may (or may not) be true of the exact sciences, we don't know without a random sampling of episodes in the history of science. (I can think of three big, glaring exceptions in psychology: Watson, Hull, and Freud.)

To answer the prescriptive question, one can proceed individually or competitively. Given an historical sampling of theories, divided into "textbook-sure" and "rejected" (for 50 years or more), we compute a running index of *percent forecasted* for each theory's successful derivations, over time. Even this crude measure should clearly reveal two clusterings of the graphs (one set rising higher over time, the others remaining lower) if the predictivists are right about the principle's epistemic validity. The competitional case is perhaps more complicated, requiring attention to countervailing properties in a discriminant function, although I am not clear about that at present.

How the ⌜*ceteris paribus*⌝ translates when we do not deal with the idealized pure case of Footnote 24 but rather have to evaluate the predictivism principle over a batch of actual theories, it is hard to say rigorously. One (crude but not useless) approach is to "match" theories in a domain (e.g., biochemistry, mammalian learning) roughly by a combination of objective properties and quantified raters' judgments, the members of a matched pair differing in long-term fate (ensconced in textbooks/definitively dead). Such matching does not require equating each duo "in all respects," an impossible demand which life science researchers never attempt (e.g., in medical studies comparing two drugs' efficacy). One matches for variables known, or feared, to exert big effects, the statistical purpose being to enhance the experimental sensitivity to a drug difference by reducing the variance of residual (intrapair) differences in a matched-pair *t* test (Fisher, 1970, 1971). The statistical question put is, "Do the long-term survivor theories differ from the discarded theories in the proportion of correct derivations that were forecast?"

A more rigorous and illuminating approach is to say, "*Ceteris paribus* is always false as between any two theories, because there are *several dozen* respects (formal + conceptual + empirical) in which actual theories differ. If one asks whether forecasted facts weigh more than nonforecasted, this question is most fruitfully construed as concerning their comparative contribution to increments in a verisimilitude

predicting function $\hat{V} = \Phi(x_1, x_2, \ldots x_m)$, hence, for number of forecasted facts x_F, to evaluate the partial derivative $\partial\Phi/\partial x_F$. If the *V*-function is linear, this derivative is a constant, a -weight in a regression equation or discriminant function. If nonlinear, the derivative depends at least on x_F itself. If *configurated*, it depends jointly on x_F itself and one or more of the other *x*s (i.e., for some other property x_j, we have $\partial^2\Phi/\partial x_F \partial x_j \neq 0$ somewhere in the realized region [Meehl, 1954/1996, pp. 132-135]). The nonvanishing second-order mixed partial derivative in the continuous case corresponds to a significant first-order Fisherian interaction term $\left(\overline{y}_F - \overline{y}_{\overline{F}}\right)_P - \left(\overline{y}_F - \overline{y}_{\overline{F}}\right)_{\overline{P}}$, where F denotes *forecast* and P denotes *precise*, these being crudely classified discontinuous *levels*, as in agronomy. One could even examine a large class of single experiments that appeared early in various theories' life histories, categorizing each experimental result dichotomously as $F/\overline{F}$ and $P/\overline{P}$, the 'output' variable *y* being ratings (or proportions) of long-term success. If the heterogeneity here involved is counterintuitive, we remind ourselves that the cliometric metatheorist proceeds in the same manner as the life insurance actuary. These are all empirical questions put to the collection of historical episodes." In the Einstein case, one suspects that the weights given light-bending and Mercury anomaly were not constants but functions of numerical precision (at least), and probably other properties as well (e.g., "theoretical depth" of derivation, as Brush points out and Eddington admitted). Of course the metatheorist will deal throughout with two such *V*-functions,

$$\hat{V}_S = \Phi\left(x_1, x_2, \ldots x_m\right)$$

the function describing the scientific community's behavior, and

$$\hat{V}_C = \Psi\left(x_1, x_2, \ldots x_m\right)$$

maximizing long-term validity (i.e., a prescriptive function). One of the descriptive tasks of actuarial metatheory is to analyze how these functions differ. There is not the slightest doubt that they will differ very considerably.

Relation between Cliometrics of Theory and Metatheory

This paper is primarily about the need for an explicit actuarial approach to conducting metatheory and is directed at psychologists, sociologists, philosophers, and historians of science who are engaged in formulating and appraising metatheoretical principles. How is this related to the actuarial study of "ordinary" (first-level, non-meta) scientific theories? The short answer is that the arguments (empirical and *a priori*) in favor of examining ordinary theories cliometrically (Faust, 1984) should apply, *mutatis mutandis*, to the appraisal of metatheory, once we have accepted the view that metatheory is itself the scientific (empirical and formal) theory of scientific theorizing. For example, we may confidently expect that the sources of erroneous cognition listed in Table 1

impair the performance of any human thinker whether the task is to diagnose a mental patient, appraise a theory about liver functions, or decide for or against Popper's metatheoretical emphasis on falsification. Of course this common rationale for proceeding actuarially does not preclude the existence of interesting and important differences between these three enterprises. The roles of values, costs, utilities, risks, etc. in clinical work will be greater than in the other two domains where we are, as Levi (1967) puts it, "gambling with truth." The percent of discourse employing metalanguage predicates (e.g., ⌜true⌝, ⌜invalid⌝, ⌜biased⌝, ⌜well supported⌝, ⌜independent evidence⌝, ⌜theoretical⌝) will be greater in conducting a metatheoretical inquiry.

Besides the obvious relation, that empirical metatheory is itself subsumed under generic ⌜theory⌝, a more interesting point is the impossibility of cliometric research on either being conducted without being forced willy-nilly into doing the other. Suppose I investigate the strategy of "Lakatosian defense" (Meehl, 1990a) in my role as metatheorist. Rejecting the unassisted case study in favor of a random sampling of scientific episodes, I have to ascertain the historical facts per episode and compute statistics on (a) what scientists holding a theory did when confronted with apparent falsifiers, (b) whether what they did "worked" in the long run, (c) what properties and relations of subsets of theories were correlated with success/failure of the strategy. All of this is the cliometrics of *theories*, albeit done for the purpose of appraising a *metatheoretical* principle [roughly, "*Ceteris paribus*, it is rational to conduct a Lakatosian retreat when confronted with falsifiers, *provided that* the theory has lots of money in the bank, acquired by having successfully predicted several Salmonean 'damn strange coincidences'" (Meehl, 1990a)]. We see from such examples that the way to research a metatheoretical principle cliometrically is to run statistics on samples of first-level theories.

The same unavoidable move goes in the other direction. If I obtain statistics about the properties and relations of theories, such statistics will include their long-term track records, including current inclusion in all textbooks as "firmly established *fact*" (*pace* the logician!) or as "totally discredited long ago" (e.g., caloric, phlogiston). Whether such dichotomized success/failure categorization is taken as quasi-gold standard criterion of verisimilitude (as by realists) or as merely an extrapolation from short-run predictive success to a longer run (as by fictionists and instrumentalists) hardly matters. Either way, we have empirical generalizations concerning correlates of "success." But this *descriptive* knowledge entails an immediate *prescriptive* move, via the straightforward hypothetical imperative, "If you aim at scientific success, you should do so-and-so, since the history of science shows that this strategy tends to succeed." This advice can be bolstered by further

metalinguistic comments *explaining why* the recommended strategy tends to succeed, the reasoning including components of logic, mathematics (especially probability theory), and old-fashioned "armchair epistemology."

As to these theoretical understandings of an actuarially derived principle, they will not only satisfy the metatheorist's *n Cognizance* (Murray, 1938) but will come to serve a useful function for the working scientist (who may have little or no *intrinsic* interest in metatheory as such). Scientists routinely invoke metatheoretical principles in appraising theories whether they label their doing it "philosophy of science" or not, especially in controversy or in periods of theoretical crisis. Most scientists metatheorize—propound, defend, or argue against metatheoretical principles—at least intermittently. (Speaking for my own field, psychopathology, I assert without fear of refutation that they usually do it badly.) It is my belief that when a sizable and coherent body of cliometric metatheory is developed and strongly tested, its principles will become part of every competent scientist's everyday toolkit, just as the concepts of statistics, curve fitting, calibration, reliable instrumentation, replication, sample bias, probability, standard error, nuisance variables, spurious correlation, control groups, randomization, etc., are today. To quote an optimistic passage from an earlier work on this subject,

> So let me wax even braver and play the prophet. I predict that the scientists of tomorrow will employ an armamentarium of quantitative indices of theory properties, as adjunctive to judgment and sometimes controlling it. It will seem quite natural to them, and they will look back on our evaluative practices with pity, wondering "How could those poor people do as well as they did in appraising theories, given the crude, subjective, impressionistic way they went about it?" (Meehl, 1990a, p. 179)

The formal components of metatheory (logic, mathematics, probability theory) will presumably be capable of high confidence predictive functions when actuarial data are incomplete, as regularly occurs in other sciences when sufficiently advanced. Thus, theorems of the probability calculus often permit one to set safe bounds on an unknown quantity provided that other values can be reasonably bounded, whether on theoretical or direct observational grounds. *Example*: Suppose large-scale metatheoretical cliometrics shows that the proportion of theories in biochemistry having property Q_1 is p_1; and a similar survey shows property Q_2 occurs with probability p_2; but these two studies were unfortunately not conducted on the same episodes (a situation that occurs continually in psychopathology research and is one reason for the current emphasis on cross-clinic, nationally organized research projects). We have no direct actuarial data base for answering the question, "What probability $p_{(1\vee2)}$ attaches to the class of theories that exhibit properties Q_1 or Q_2 *or both*?", the properties being nonexclusive. Under some such circumstances, a plausibility argument can be made to treat the studies as (approximately) samples from the same domain, and we can then rely on the theorem

$p_{(1\vee 2)} = p_1 + p_2 - p_{12}$ if we can set reasonable bounds on the value of p_{12} despite not knowing it with any precision. A fairly small sample from a closely similar domain, perhaps accompanied by purely analytical considerations about the internal structure of theories, might suffice to assure us that not more than, say, 1 theory in 10 combines properties Q_1 and Q_2. Hence if our large-scale actuarial estimates were $p_1 = .38$ and $p_2 = .51$, we may write

$$p_{(1\vee 2)} \geq .38 + .51 - .10 = .79 .$$

If in addition we have extensive actuarial data as to the success-rates associated with properties Q_1 and Q_2, and *absent analytical considerations suggesting that Q_1 and Q_2 somehow work adversely when conjoined*, we can obtain a lower bound on domain success frequency. (Note that this sort of argument is not available to a metatheorist who excludes philosophical analysis from the conceptual toolkit.)

PSYCHOLOGIST AND LOGICIAN: AN EXAMPLE OF HOW COMPLEX THEIR RELATIONSHIP MAY BECOME

When we understand that (1) metatheory is the empirical, history-based scientific theory of scientific theorizing and (2) this definition does not forbid, but rather requires, that the traditional "philosopher's" tools of logic, probability theory, and armchair epistemology be employed in fulfilling the *explanatory* function of metatheory, then how the enterprises labeled "psychology" and "philosophy" are to be integrated is of first-rate importance. In this section, while I shall make a few assertions and denials, I mainly want to exemplify the complexity and difficulty of the integrative problem, hoping to induce members of both professions to work on it. I shall ask more questions than proffer answers. The italicized word ⌜*Query*⌝ beginning a question sentence is to be taken literally; the question put is not rhetorical or as a minor curiosity in passing. It signals—to logician, historian of science, psychologist, or all three—that what follows is an important problem, deserving our intensive interdisciplinary efforts.

A major component of cliometric metatheory—I believe the most important—is to *ascertain* and *explain* the correlation between theories' verisimilitude and their "track record" (success in deriving facts). I have offered a proof (Meehl, 1990b) that even a crude index of verisimilitude (how many postulates in Omniscient Jones's true theory T_{OJ} have been deleted or altered in T) will correlate highly (over theories) with an equally crude index of their factual performance (how many experiments are successful, a yes-or-no classification) in the long run. As of this writing, no one among a couple of hundred philosophers and psychologists who received this technical report has communicated an objection, given my bypassing of Hume's problem.[29] But my proof ignored differences

[29]Roughly, Hume's problem is the lack of a noncircular argument for *any* inductive inference from past empirical regularities to future ones. A closely related "Hume problem'—some would equate them—is how the notion of *nomic necessity*, the difference between a natural law and a so-called "accidental universal" statement, is to be explicated and justified.

in the *a priori riskiness* of the factual prediction, which working scientists weight heavily in appraising theories. The proposed corroboration index *C** presented briefly above (in more detail in Meehl, 1990a) is a rough first step in numerifying the riskiness feature. It differs from Popperian "risk" in an important respect, namely, a *near-hit* counts in a theory's favor when the predicted value allowed is a small region of the antecedent Spielraum. This seems to accord with scientific practice, which usually counts a near-hit in a theory's favor, despite its being a Popperian falsifier *modus tollens*. A Popperian could handle this paradox easily, by distinguishing *falsifying T* from *totally abandoning T*. Since theories in psychology are almost all literally false anyway, the difference between concluding "*T as stated* is literally false" and "*Since T as stated* is literally false, we will simply junk it and start afresh with some new, totally dissimilar set of ideas" is a crucial difference in research policy. [Of course there are additional reasons for continuing to work on an apparently falsified *T*, such as the auxiliary theories being somewhat problematic, and the *ceteris paribus* clause highly problematic (Meehl, 1990a, 1990b).]

I have labeled as *Salmon's principle* the claim that theories acquire "money in the bank" (a rational claim on credence, hence warranting defense despite falsifiers) by successfully predicting facts that, absent the theory, would be "damn strange coincidences," as Salmon says (see Meehl, 1990a). For ordinary prediction of an observational numerical value, this strange coincidence consists of the theory tolerating a numerical range that is a small portion of the Spielraum and getting it right *or nearly right*. The principle is suggested by—I do not say deduced from—a broad piece of advice, "If you want to achieve a causal understanding of the world, do not adopt a policy of attributing replicable orderliness to mere coincidence." This is a weak and negative principle, since it offers no assurance of success. But it is surely sound, because a *policy* of accounting for observed orderly relations as "coincidence" would *mean* that we avoid concocting explanatory theories. So if there *are* any true explanatory theories, and if humans *can*, sometimes, come by them; we won't succeed at that if we proceed on the "merely a damn strange coincidence" principle. The reasoning here is similar to Reichenbach's (1938) justification of the inductive straight rule, and—from a noninductivist—Popper's epigraph quoting Novalis, "Theories are nets; he who will catch must cast" (Popper, 1935/1959). Here, as in my previous papers on this topic (Meehl, 1990b), I allow myself the luxury of bypassing Hume, not because Hume's problem is "pseudo" or of no philosophical interest, but because (1) I meet no working scientists who worry about it and (2) I am firmly convinced, like Popper and others,

that it cannot be solved. I share the view of Grover Maxwell (1974), following Bertrand Russell (1948; and see Hawthorne, 1989), that the logical positivist ideal of reconstructing scientific knowledge with *no* apparatus except logic and mathematics—no metaphysical presuppositions, simon-pure *statement empiricism*—cannot be realized. I therefore postulate (Latin *postulare*, to demand) that the "external world" exists, it is objectively "out there," would be there whether I knew of it or not, and further, that it is composed of more or less *permanent kinds* which follow *natural laws*. I take the aim of theoretical science to find out about those entities and their laws. (I could even argue—having bypassed Hume—on "inductive evidence" to date, that the human mind is so constituted as to be sometimes *capable* of ascertaining these matters.)

Salmon's principle is qualitative, as is the underlying advice. Adopting a policy to try explaining observations otherwise than as "mere coincidence" is rational, but as it stands does not provide *degrees* of evidentiary support. We need to assign such degrees, because we give credence to some theories more than others, even in falsification. Auxiliary theories are affirmatively *trusted* when we rely on them as part of a *modus tollens* falsification, hence a less-trusted auxiliary means a weaker discorroboration of the main theory. Technological "trusting" of one theory more than another is unavoidable when life and death, millions of tax dollars, etc., are often at stake. In the context of discovery, the scientist does not normally decide whimsically whether to pursue a theory, attempt to refute or amend one, engage in puzzle solving, concoct a new theory, or whatever. One may be content to invoke Santayana's "animal faith" (1923) as to Hume's basic problem—that doesn't bother me, or the psychologists I know—but, *after bypassing Hume*, we do not say that specific theory-appraisals are also sheer faith, devoid of any rationale, immune to criticism or argument. Therefore we would prefer to have a proof, at least a "plausibility argument," that when a theory successfully predicts facts that, absent the theory, would be a damn strange coincidence, it probably has some verisimilitude. A skeptic, *without* relying on Hume's rock bottom epistemology, could say, "Your theory *T* predicted a numerical value in a narrow range of the Spielraum; it had a high intolerance, and it predicted correctly. As you say, a damn strange coincidence. So what? Why should that impel me to accept your theory?" Or, less skeptical but still pressing for the argument to be spelled out, "Your theory and mine both predict the observed fact, neither is falsified. Yours was much more intolerant, took a bigger risk, passed a more severe test, reveals a damn stranger coincidence. So what? How does that render yours more deserving of credence than mine?" A metatheoretical account of theory appraisal that cannot answer this skeptic, explaining the rationale of such a basic feature of scientific method, can hardly be considered satisfactory.

(A similar question has been raised about statistical significance testing: The *t* test passes significance level $\alpha = .01$. "Such a difference would occur in less than 1 in 100 samples by chance, if there were no population difference." "So? Why shouldn't that have happened on this occasion?")

What we would like is a probability, going *backward* from the facts, instead of an *im*probability, going forward from "no theory" *to* the facts. We would like to say something like, "Since my predicting the observed value, absent theory, so *narrowly and accurately* is *a priori* improbable—picking out a small region of the antecedent Spielraum and getting it correct—I infer that the theory is more probable than theories that mispredict or predict correctly but with larger tolerance, with *less* of a damn strange coincidence." Many hold that no such inference is attainable. They may be right—I often think so myself—but let us have a try at it.

Among the various meanings of the term ⌜probability⌝—one recent author whose name I cannot recall managed to list 7—the one we need here is the one most familiar to psychologists, a *relative frequency*, Carnap's probability$_2$ discussed above. Whether, in our problem, we can *infer* an empirical relative frequency from knowing a set of categories, as the principle of indifference (or principle of insufficient reason) is used to do in the classical probability calculus of *a priori* "equally likely cases" in games of chance, I shall consider later. If so, it will be necessary to state the empirical conditions legitimating such a move—whatever in human theory-making would correspond to physical stipulation like "The roulette wheel is unbiased," leading via a Spielraum argument to $p = 1/37$ for a single number win at Monaco ($p = 1/38$ at Las Vegas because of the 00). Frequentists have no objection to this use of the principle of indifference, witness R. A. Fisher's rationale for random number tables in the design of experiments.

To obtain a relative frequency, we must specify a reference class. What is the reference class of theories? If the reference class of theories is *infinite*, as some have alleged (on purely "logical" grounds), it seems generally agreed that no kind or amount of evidence could make the probability of a particular theory nonzero or at least noninfinitesimal. This is one standard objection to probabilifying theories via Bayes's Theorem, because whatever may be the value of the product $p(T_i) \cdot p(\mathrm{e} \mid T_i)$ = Prior probability of theory × conditional probability of the evidence on the theory, the sum $\sum_{\infty} p(T_j) \cdot p(\mathrm{e} \mid T_j)$ of all the analogous products for the infinite set of competing theories will swamp the one of interest in the denominator. (Since the probability calculus, on anyone's axiomatization, *defines* a probability as a real number lying in the interval [0,1], this argument would seem to force an infinite set of the priors to infinitesimals to prevent the denominator sum from exceeding unity.) Popper does not believe that corroboration (or Carnap's "confirmation") can even *be* a probability, claiming to have proved that it

does not satisfy the axioms of the probability calculus, a claim that Carnap and his followers have never conceded. (To a nonlogician like myself, this seems odd, since one would think that such a "formal" question would have a clear-cut rigorous answer. I shall not attempt to discuss it further.) [*Query*: "If the class of logically possible theories to derive a given set of facts is infinite, could it nevertheless be the case that some proper subsets, appropriately defined, would have larger probabilities, collectively, than others?" I suspect this query takes us into technical matters of measure theory which I lack the mathematical competence to discuss.] But if the reference class of scientific theories (about a given empirical domain) is *finite*, however large, it may be possible to link (stochastically) the relative frequency truth to the antecedent improbability of their successful "damn strange coincidence" predictions.

In what follows I shall refer to the *truth*-frequency of a class of theories, despite our recognition that almost all actual theories are literally false. Not *all*, contrary to a common assertion. It is literally true—no idealization or approximation—that we think with our brains, that glomeruli filter nitrogenous wastes, that the liver stores glycogen and secretes bile. It is literally true that the genetic code lies in cistrons, which are ordered sequences of codons, which are ordered triplets of the bases adenine, guanine, cytosine, and thymine. It is literally true that when a radium atom's nucleus undergoes radioactive decay, the three kinds of "rays" emitted (α, β, γ) are helium nuclei, electrons, and electromagnetic waves shorter than "hard" X-rays, respectively. It is literally true that the earth revolves in a quasi-elliptic orbit around the sun, at a mean distance of between 90 and 95 million miles. (Note how we can assure literal truth even of quantitative theories by the simple device of substituting numerical tolerances for imprecise point values.) For theories that are *conceptually idealized* or *quantitatively approximate*, we achieve the literal truth condition by appropriate qualifying comments in the meta-language. Before van der Waal's correction to the gas law, physicists already took it for granted that the two idealizations of molecules as (a) point-masses (hence nonspace-filling) and (b) having no mutual attractive forces, were literally false. The van der Waals revised formula itself relies on that metalinguistic recognition; but could we not say that prior to his derivation the [theory + metalanguage] composite could fairly be labeled true? Thirdly, I shall follow Grover Maxwell in counting a theory as true when it has a very high verisimilitude. It would be cumbrous to use this phrase (and hard to prefix to "-frequency"), and I shall carefully avoid making any illegitimate use of this approximative convention. So ⌜true⌝ in what follows covers

1. Literal truth, as it stands,

2. Literal truth, satisfied by metalanguage warning about the object-language idealization,
3. High verisimilitude.

I must emphasize (especially for a readers lacking philosophical sophistication) that ⌜true⌝ is an *ontological* (not an *epistemological*) concept. It characterizes propositions that are objectively correct, that assert what is the case in the real external world whether we know it or not. It is not a concept about evidence, grounds, rational belief, observational support. These latter are evidentiary matters, and we must not conflate the *truth* of a statement with purported *proof* of it.[30]

In considering the truth frequency of empirical theories in a domain, I shall use the phrase ⌜postulate sets⌝ despite the positivist-bashers poking fun at the idea. Whether or not a scientist bothers to list in any formal way the basic conjectures of his theory, from which he intends to derive the factual observations of the domain, it is not hard to ascertain what they are. The easiest way is to ask him and, lacking that, to inquire how he proceeds in deriving experimental results, how he talks about possible amendments; and then of course we can objectively ascertain from a system of related propositions—whatever the scientist *says* to the contrary—what certain internal, logical, and semantical relations are. For example, it is quite possible that the scientist mistakenly supposes that he needs a certain postulate when it can be shown to be derivable from the others; this has frequently happened even in purely formal systems such as Whitehead and Russell's *Principia Mathematica*. Or perhaps he thinks that he can derive a certain conclusion when in fact the derivation does not go through. With the exception of the set theoretic

[30]The worst error of Vienna positivism was Schlick's "the meaning of a sentence is its method of verification," which was quickly amended, qualified, and finally abandoned. The truth of the statement ⌜Caesar crossed the Rubicon⌝ depends on one and only one fact, necessary and sufficient for its truth, namely, that he crossed. A Roman centurion, his aide-de-camp, verified this statement by direct observation and, later, memory. For us today, the verification (it is only the weaker *confirmation*) consists chiefly of some "mounds of ink" (Neurath's felicitous term) on a 9th century palimpsest in the Vatican library—our earliest manuscript of *De bello Gallico*. If Schlick's dictum were accepted, the centurion and I could not be believing (even roughly or essentially) the same thing, since our methods of verification are totally disjunct! The empirical content of "Caesar crossed the Rubicon" is identical with that of "The statement ⌜Caesar crossed the Rubicon⌝ is true," and one who holds there is adequate evidence for asserting the first need not be shy about asserting the second. The two differ merely in that the former is object language, the latter is metalanguage. Needless epistemic timidity about ⌜true⌝ was taken care of over a half-century ago, by logician Alfred Tarski. (Knowledge sometimes diffuses slowly!) Many psychologists suffer from "truth-phobia," probably because they were taught a simplistic operationist-verificationist theory of meaning—similar to Schlick's blooper, conflating proof and truth—in a behind-the-times introductory psychology class. On that view, allowing oneself the (innocuous but necessary) meta-term ⌜true⌝ sounds rather like claiming "absolutely proved," "certain," "not to be questioned." Such certitude is forbidden as a kind of medieval, transempirical dogmatism, not in the spirit of modern science. Readers who can introspect this difficulty, stemming from a confusion of truth with proof, may find a quick and painless cure in Popper's "Truth, rationality and the growth of scientific knowledge" (1962, Chap. 10, especially pp. 223-231), Popper (1972, pp. 319-335), Carnap (1949), and Tarski (1935/1956, 1944).

model approach of some recent philosophers of science (beyond the scope of this paper), those who poke fun at the logical empiricists for talking about theoretical postulates almost never tell us what they have in mind instead when they allege that a theory "explains," or "predicts," or "accounts for" observational facts. Already when I was a senior in college hearing lectures by logical positivist Herbert Feigl, I was puzzled by the notion of a "probability implication," except when it referred to theories containing numbers or quantitative relationships in the postulates as a result of which (as in dealing with gambling problems) the probability calculus could be used to derive certain numbers from others. My present view, which I shall not try to defend here, is that whenever you find a metatheorist talking about a "probability implication" (or any equivalent expression) and it is not a matter of deriving statistical frequencies or reasonable bets about a particular from certain given numerical assumptions operated upon with the formalism of the probability calculus, it will turn out that a deductive relation holds between the premise and the conclusion, provided that certain auxiliary assumptions are considered nonproblematic and—of the greatest importance—that there is a *ceteris paribus* clause. It is as if one were saying "If these auxiliaries are taken as correct, and if the *ceteris paribus* clause is correct (so that no other unmentioned but systematically influential factor is operating), then it would follow deductively from theory *T* that so-and-so happens in the lab or the clinic file statistics." What confuses this kind of situation (which is extremely common in the biological and social sciences but also arises in astronomy, physics, chemistry, and geology) is that, to make the deduction go through strictly, we need both the presupposed auxiliaries and the *ceteris paribus* clause, and we know that they are not for sure. Therefore, I think the idea of a "probabilistic implication" can in all reasonable cases be reduced either to the application of the probability calculus to relative frequencies, known or conjectured on the basis of the principle of indifference; or, for cases not fitting that model, it amounts to a valid deduction but with necessary components that one knows and is willing to explain in the metalanguage in talking about the theory may not be correct. Feigl used to speak of a "theory sketch," the scientist's statement of an avowedly incomplete and somewhat vague theory, with numerous "promissory notes" as to hoped-for development, and he warned against rejecting such "theories" on perfectionistic grounds. But even those theory sketches, I think, are defended by "sketching out" deductive derivations, with a metalanguage comment about the approximations, idealizations, tentative auxiliary conjectures, and *ceteris paribus* clause required for the derivation to go through.

What is it that empirical science demands of the postulate set constituting a theory? "To explain the facts" is the short way of saying it. We therefore start with a set of observational statements which are well formed(*wff*s). For some analytic purposes,

one considers the class of observational *wff*s accepted by the scientist appraising a theory at a given point in time; for others, we may conceptualize the set of all such *wff*s that will be accepted into the belief corpus of the scientific community, based upon all experiments reported before the sun burns out; and finally it is sometimes useful to conceptualize the set of all *wff*s that would have been empirically confirmed if the experiment had been performed, although it was not.

The postulates that compose the theory are of two kinds, those that contain only theoretical terms, and those that contain both theoretical and observational terms. In order for the theory to be an empirical theory (rather than metaphysics) it is necessary that the theoretical concepts be hooked somehow with observations, and this is done by means of what are sometimes called *bridge laws*, other times *operational definitions*, and were called by Carnap *meaning postulates*. I bypass here the interesting question whether some of these bridge laws have the character of definitions and others of assertions, or perhaps perform both functions at the same time. Feyerabend (personal communication, 1959) argued that all bridge laws ultimately should or will become themselves theorems. For example, the bridge law relating the height of a mercury column in a thermometer to the theoretical concept of temperature defined as the mean kinetic energy of the molecules is a theorem of statistical thermodynamics: the impact of the molecules in the hot soup impinging upon the glass of the thermometer increases their motion and this increases the motion of the molecules of the mercury and forces the column to expand. I am inclined to agree with Feyerabend about this with regard to utopian science, with the single exception of the mind/body bridge law. When the neurons in my Brodmann area 17 are firing in a certain way, I experience a raw feel of green quality. I am inclined to agree with Du Bois-Reymond, that this psychophysical nexus will forever remain incapable of scientific explanation.[31]

We can conceive of variously defined collections of postulate sets, each including its successor subset as follows: First we have postulate sets that are *logically possible* as explainers of the class of observational *wff*s. Philosophers of science regularly state that this set is infinite and show surprise when asked who proved that theorem. One might expect that such an important theorem of metalogic would be named after somebody, like the Goedel Theorem, Church's Theorem, the Loewenheim-Skolem Theorem, etc. If one

[31]Emil Du Bois-Reymond enumerated seven deep mysteries of the universe, four of which he considered insoluble and three soluble in principle: the nature of matter and force (insoluble), the origin of motion (insoluble), the origin of life (soluble), the apparent purposefulness of nature (soluble), the origin of sensory perception (insoluble), the origin of reasoning and language (soluble), and the freedom of the will (insoluble). I am indebted to Professor Martin Carrier for help in tracking down the complete list (personal communication, December 1991), but neither he nor I know of an English translation of the list. One German source is Du Bois-Reymond (1880).

persists in asking how we know it's true, the usual response is to show how a conjunction of statements that has an implication relation to another set can be altered by adjoining various conjunctions and disjunctions, which trivializes the whole matter. I do not accept any such examples as a satisfactory demonstration because no empirical science is analyzable with the resources of the propositional calculus, in which the smallest unit is a proposition. All empirical sciences get their deductive fertility, their technological power, and their intellectual interest by virtue of complex interconnections between the "innards" of their postulates, and even the predicate calculus is probably not an adequate way to express these kinds of semantic relationships. I am not asserting that no such theorem could be proved, but I would like to see exactly what the theorem says and what is the precise general statement of the problem conditions, so I can evaluate its importance for empirical science.

Within the class of logically possible postulate sets capable of deriving the collection of observational *wff*s in an empirical domain, not all are *methodologically admissible*. There is a set of strict rules, and a larger set of guidelines or principles, which lead scientists to classify certain kinds of postulate sets as inadmissible. What this comes down to basically is a *definition of the game of science*, as Lakatos calls it. Like chess, hockey, or charades, the game of science is defined by a system of interconnected rules. Nobody has to play the game of science, but if he purports to be doing so and persistently breaks the rules, we fault him for it and call the community of scientists' attention to it. Some of these rules are strict enough to be called "rules" but most of them, as discussed above, are only principles or rough guidelines or policies. Feyerabend's much abused (because misunderstood) "anything goes" does not mean that one policy is on the average just as good as any other or that there is no optimal way to proceed at a given stage of a particular scientific development; rather it points to the fact that there is hardly a single rule, and clearly not a single guideline or principle, that has not on occasion been violated by a scientist or a whole group of scientists with resulting scientific success. Ideally, a scientific principle should be supported by a combination of *empirical evidence* and *theoretical arguments* such as conspicuous success in the history of science, being a precept quasi-universally subscribed to by contemporary scientists and at least statistically followed by them and shown by considerations of logic and the theory of probability to be plausible means to the scientists' epistemic end (Laudan, 1977, 1984, 1990).

The question of "who is a scientist?", while it may be fun for cocktail party conversation, is not one of deep interest or difficult solution. It is answered by a mixture of cooptation and social perception by the in-group and the out-group. If Jones thinks of himself as a scientist, and he thinks that Smith is a scientist, and Smith entertains

the same two beliefs, and neither of them thinks that Pastor Fisby or Senator Claghorn or philosopher Hegel is a scientist, and the latter don't think they are scientists, but they do think that Smith and Jones are scientists, then we have the answer to our question. Nor do I find the question whether psychology is a science particularly interesting. Some parts of it are in an advanced scientific state; a great deal of it is in a fairly primitive state of scientific development both as to the replicability of its empirical findings and the plausibility and empirical support of its theories; and still other portions are hardly scientific at all. One must, of course, distinguish between whether the scientific method is being applied although with rather poor success due to intrinsic difficulties (Meehl, 1978), stupidity, or bad luck, and whether the scientific method is being pseudo-practiced (Feynman's "cargo-cult science," Lykken, 1991; Feynman, 1986; Andreski, 1972), or whether no pretense is even being made to proceed scientifically. If these distinctions are not made, a great deal of confusion results.

From the (allegedly infinite) set of alternative postulate sets "logically (formally)" capable of deriving a specified finite set of observational *wff*s, some are deemed *methodologically inadmissible*. I have offered a short list of some kinds of methodological inadmissibility (Meehl, 1990b) and will not repeat it here but will only give some illustrative examples. (My list of eight methodological proscriptions was concocted in not more than a half-hour of casual reflection on a stroll; a statistical sampling of the critical literature of various sciences, and of metatheoretical writings, plus some spade work by statisticians and logicians, should be capable of enlarging my list many fold.) *Example*: Considering "input-output" theories such as we find in the life sciences and social sciences, suppose the theory postulates a *pure intervening variable* (*IV*) θ (MacCorquodale & Meehl, 1948) mediating the observable relationship between an input variable x and an output variable y, thus $\theta_1 = f_1(x_1)$, $y_1 = g_1(\theta_1)$. And suppose that the theory has another pure IV linkage $\theta_2 = f_2(x_2)$, $y_2 = g_2(\theta_2)$, and so on through a set of such. There are no cross-connections between the θs, and there are no "output forks" in which a single θ controls two or more output variables. This so-called "theory" consists simply of a heap of postulates linking nondescript, not further interpreted, and unconnected intervening variables to single inputs and outputs. A postulate of this kind I call an *isolate*, and it is obvious that in no field of science would a theory be considered acceptable were it composed of nothing but a collection of such parallel, disconnected isolates. *Example*: Suppose in such a two-link postulate the output link is the mathematical inverse of the input link, as if we were to say that a certain inner psychological state θ varies as the square of a quantitative stimulus property x, and a quantitative response property y is proportional to the square root of the inner state θ. We have $\theta = \mathrm{A}x^2$, $y = \mathrm{B}\theta^{1/2}$,

hence input-output relation $y = Kx$. Such a *formal undoing* is also proscribed by scientific method. In saying this, I am assuming that nothing further is conjectured theoretically or planned empirically to connect the isolates or to rationalize the formal undoing. Suppose a theorist offers us what Herbert Feigl called a "promissory note," such as that he plans to conjecture some cross-connections between the isolates; or he has some plausibility arguments, perhaps based upon other knowledge of the kind of system being studied, suggesting a microprocess in which θ would be expected to go up with the square of x, and another microprocess in which y would increase with the square root of θ. Given such a sketch of his future theoretical or experimental developments, we may choose to be patient and await further developments. But this policy of permissiveness in the context of discovery does not contravene the general principle that a scientific theory *as it stands* is not acceptable if it is merely a heap of isolates, or if the input-output connections involve nondescript intervening variables in which the output mathematical functions simply reverse the effect of the input function. *Example*: We do not countenance theories in which the list of theoretical predicates is longer than the list of observational predicates. (Of course the theory may postulate numbers of individual *entities*, characterized by the predicate list, which are vastly in excess of the observational entities, as when we know that there are many more electrons in a chair or table or stone than there are tables and chairs and stones all together.) This insistence that explanatory properties should be fewer in number than the properties and relations to be explained at the observational level is commonly attributed to a liking for parsimony, but since I do not accept the law of parsimony as usually understood,[32] I will not try to defend the principle but simply point to it as one that scientists universally accept. *Example*: Certain kinds of mathematical legerdemain are forbidden. Suppose the input-output relationship is of the form $y = \log x$, and we express the input-theoretical function as the infinite series $\theta = \log x = (x - 1) - \frac{1}{2}(x - 1)^2 + \frac{1}{3}(x - 1)^3 - \ldots$. Suppose the attainable precision of observational measurement is such that we can be confident, at least in the foreseeable future and maybe forever, that, given the size of the units (for instance as in a psychometric test where the scores move by integer steps of one item), the remainder

[32]"Prefer the simplest explanation" has been severely criticized by philosophers on several counts: ⌜Simplicity⌝ is hard to define for this purpose, except in (some!) curve-fitting decisions. Different definitions of the term, each intuitively appealing, can conflict with respect to a given theory. No ontological metaproof exists that the world must be "simple," or that "simpler" theories are, as a class, more likely to be correct. The most impressive theories of mature sciences are hardly ever "simple," either conceptually or in the formalism. The only four kinds of simplicity I accept as relevant—not dispositive—are (1) curve-fitting, (2) computational ease (technological), (3) stronger falsifiability (Popper), and (4) no theory concocted to explain facts already explained by an existing corroborated theory (Ockham's *entia non sunt multiplicanda praeter necessitatem*, which I am told he never stated in that form).

term in the series stopping after 145 terms would not be empirically discernible as error. We write each of the terms as an intervening variable $\theta_n = (1/n)(x - 1)^n$, and then add postulate #146 which says that output variable $y = \Sigma\theta_n$. We add an interpretive text to the operational text in which we speak of each of these θs as being the numerical value of some intervening variable not further characterized. Note that this is a perfectly good postulate set in the sense of the logicians' claim that an infinite number of postulate sets will explain the observational fact that $y = \log x$, but of course it is totally preposterous as a scientific theory and no sane scientist would even contemplate such a thing.

Substantively inadmissible concepts constitute an important class of exclusions, although one has the impression that these constraints are tacitly understood, presupposed without explicit mention, probably because they are less frequently violated than the methodological sort. I can discern two main categories of substantiative prohibitions, with some overlap. Whether there are additional categories, I do not know. The first is hard to explain because it forbids certain *kinds of entities* and *forms of relation* that I cannot characterize generically. Entities possessing consciousness, cognition, and purpose are automatically excluded as explanatory conjectures except when the subject matter is living creatures sufficiently developed so that such processes are not out of the question for them. Post-Galilean scientists, and most educated nonscientists as well, take this proscription so much for granted that we hardly realize what a step forward it was in the history of human thought. If, for example, something should seem amiss with the theory that milk is curdled by enzymes secreted by *Lactobacillus bulgaricus*, biochemists and others would begin casting about for an alternative theory, but no one would countenance reviving the medieval housewives' theory that milk is curdled by the malignant activity of brownies. Ghosts, gods, spooks, vital forces, karma, entelechies, diathetes (Kapp, 1940, 1951), élan vital, angels, demons, djinns—all are "out" as explanations from the scientific point of view! For some behaviorists, "mental" events are excluded.

Even if the entities spoken of are not forms of animism, explanation in terms of certain kinds of properties and relations are also excluded. Thus Copernicus, despite making the sun the center of the solar system, was still saddled with epicycles (although fewer of them than Ptolemy was) because he required circular orbits on the ground that the circle was the "perfect figure." Aristotle conceived of gravitational effects in terms of the "natural end of a body." The medieval cosmology was permeated with evaluative notions such as the idea that things beneath the lunar sphere were "corruptible" while supralunar entities were incorruptible. The alchemists classified certain metals as

"noble." While there are some *apparently* teleological notions in physics, such as least action principles, as I understand it these are themselves, as applied to any concrete situation, derivable, although computationally they may be much more convenient than formulations in terms of a micro-understanding. There is a tendency among the positivist bashers to emphasize so-called "aesthetic" considerations in physics, but I find such talk usually misleading and tendentious, because notions of symmetry, beauty, and *depth* turn out to involve a cognitive appreciation of complex structural relations in the formalism, a rather different thing from expressing one's aesthetic appreciation of the Venus de Milo!

The rejection of purposive, teleological explanations by post-Galilean science is so strong as to lead to a purpose phobia even when studying organisms that quite clearly have purposes, such as the goal seeking of human beings. Thus we find Tolman in his classic *Purposive Behavior in Animals and Men* (1932) writing at length apologetically as to why it is all right to introduce purpose, to such an extent that instead of saying (correctly) that purpose is inferred from the behavior, he tries to make out that it is "immanent in" the behavior itself.

In addition to the substantive constraints excluding certain kinds of entities, processes, and relations of an evaluative, mentalistic, and teleological sort when these are not being attributed to goal seeking, living organisms, a very powerful set of constraints exists in any science by virtue of its location in Comte's pyramid. While Comte himself was only partly a reductionist, the history of science abounds with examples of *successful reduction* in which the concepts and laws at one level of the pyramid are derived from the concepts and laws of the next level below it, together with the necessary compositional statements. To have a generic bias against reductionism is surely a strange attitude, when many of the major breakthroughs in scientific history consist essentially of successful reductions. (What is the biggest breakthrough in the last half of our century—the Watson-Crick solution of the gene problem—but a beautiful instance of reduction?) One need not hold a metaphysical dogma against emergence—although I suspect that most psychologists subscribe to such a view whether they know it or not—to entertain a weaker form of reductionism which *forbids certain kinds of explanations* without necessarily asserting that a complete positive reduction is attainable. We would not countenance any theory of genetics and cell division which postulated that the spindle fibers in mitosis were fine platinum wires, even if that preposterous conjecture did succeed in predicting certain experimental facts. The pure logician's notion of an infinite set of alternative theories to explain the known facts is pretty much irrelevant to the way empirical science actually works when one contemplates the development of the theory of the gene culminating in the Watson-Crick discovery of 1953 (see Meehl, 1990b, pp. 24-29).

Finally, within the class of logically possible, methodologically and substantively admissible theories, we have the subclass of *psychosocially possible theories*. By that I mean not exactly "possible" (in the sense that no nomological strictly prevents them), but that there is a non-negligible disposition in the community of scientists to concoct such-and-such a theory. I assume here that dispositional properties are acceptable since there seems to be almost complete consensus among philosophers of science that dispositions cannot be avoided either in common language or in scientific theorizing.[33] We are not accustomed to talking about theories that are psychosocially possible in the strict sense that to concoct them would not violate the laws of nature or the human mind, but their probability of being conceived, at least in a given stage of scientific development, is negligibly small. Negligibly small is of course an intrinsically arbitrary concept, and I would have no objection to adopting Buffon's value $p < 10^{-4}$ as being a value so low that in our ordinary affairs we treat it as if it were zero (Meehl, 1977). Such a value would obtain for an admissible *T* if about 100 "theorists" in the community of scientists were all theorizing independently and each of them was .999999 certain *not* to conceive of *T*. If instead of Buffon's "quasi-impossible" value we choose the statistician's familiar .01 level, each of our 100 theoreticians still has a disposition of only $p = .0001$ of inventing *T*. This subset is obviously one that we cannot specify since, if we began to list such theories, we would ourselves be conceiving them. But the *concept* is empirically meaningful. If it is understandable to speak of a schizotype who does not fall ill with clinical schizophrenia, a sugar lump that is soluble but is never put into solution, a person born with a talent for being a concert violinist who never studies the violin, an election that could have been won if only a certain unfortunate speech had not been made—the examples are all around us both in science and in common life—I see nothing outlandish in considering an unknown class of scientific theories that at least one scientist had the potential (disposition, power) to invent but, for any number of causes, never got around to doing it. It may be that there are correct theories about certain domains which it is literally beyond the powers of the human intellect to conceive, just as I can understand theories that a high-grade moron cannot, the moron can understand concepts that the white rat cannot, and the rat can form concepts that the earthworm cannot. But given admissible theories that no scientist actually conceives, my point is that some of these are conceivable even though they were not in fact conceived; and among those not in fact

[33]I remember once Arthur Pap, at a meeting of the Minnesota Center for Philosophy of Science, playing devil's advocate for the notion that *all* physical properties, if carefully analyzed, are dispositional in nature, and Wilfred Sellers arguing that *almost* all of them are, but that there have to be at least a few of what he called "stuffing properties."

conceived, some had a *non-negligible probability*—wherever we set that numerical value—of being conceived, and others did not.[34]

The class of scientific theories that are logically possible, methodologically and substantively admissible, and have psychosocially a non-negligible disposition to be conceived (and hence included here are those that are in fact conceived) I shall call *accessible theories*. Among accessible theories, the subset that are actually conceived can be further divided into those that a single scientist conceives briefly (maybe in the middle of the night and forgets in the morning), those that some scientist at least writes down, then those which the scientist conveys to his colleagues and students, then those that are published in some scientific journal, and finally those which the scientific community takes seriously enough to criticize, experiment upon, and so forth.

In attempting to show, at least by a plausibility argument (which becomes a genuine proof if the plausible conditions are granted) that theories which successfully predict low probability observations—Salmonean "damn strange coincidences" of the sort that would yield a high C^* in my corroboration index—can be expected to be more frequently correct, I shall consider an extremely simple case, that of a two-postulate pure intervening variable (IV) theory in which the IV θ is initially not further characterized by any interpretive text. This nondescript pure IV is an extreme case of a so-called implicit definition via the postulated network of empirical laws, because in the cases I am about to consider, nothing is said by way of characterizing its nature other than (say, in psychology) its being physically located inside the organism, a mere "state" or "event" or "process," and that it has a *magnitude*, and this magnitude appears in two functional relationships, the input and the output functional link. Thus I am dealing here with what was forbidden above as a methodological proscription, namely, an isolate of the form $x \rightarrow \theta \rightarrow y$. [*Query*: Will the following argument generalize to admissible cases where isolates are forbidden? That is, if we have some connections between different θs and some input and output forks, is the plausibility argument for a correlation between *truth frequency* and *intolerant successful predictions* strengthened or weakened?] The theorist

[34]For an interesting imaginary example of the cognitive disposition problem, see my reply to Popper's paper on determinism (Meehl, 1970, pp. 351-354 [1991, pp. 76-78]) where I consider the case of a mathematician who *could* have found the proof of Fermat's last theorem had he worked on it a couple of weeks longer. A utopian neurophysiologist who knows about the mathematician's (micro-level) CNS dispositions is therefore able to derive that proof from analyzing those dispositions which were never realized. He takes the proof (which he, being a neurophysiologist and not a mathematician, does not understand) over to the mathematics department, where it is recognized as a valid proof. Interesting question: who proved it? It looks as if neither the mathematician who gave up too early, the physiologist who didn't understand it, nor the mathematicians who merely recognized it as valid can be said to have proved it, yet the proof came into being.

assumes both of these links are positive, for some substantive reason based on knowledge of the domain. Suppose the mathematical form of the input-output link $y = h(x)$ is logarithmic. (I bypass curve fitting problems and I assume satisfactory replication of this functional form in the lab.) With only two functions $\theta = f(x)$ and $y = g(\theta)$ available to give rise to this observational relation, we only have one possibility mathematically. The input link must be logarithmic and the output linear. (Had we allowed one of the relations to be negative, the first one $\theta = f(x)$ could be a reciprocal, and the second one $y = g(\theta)$ the integral of θ, which would give us a logarithmic input-output relation in a second way.)

Now consider the case in which the observed relationship is a decelerated power function $y = kx^c$ ($c < 1$). This we could obtain by the internal relations $\theta = ax^m$, $y = bx^n$, or by one being a power function and the other linear. Speaking in terms of acceleration and deceleration, for two power functions they could both have exponents < 1, or one unit exponent (i.e., linear) and the other < 1, or we could have one with a positive and one with a resultant negative acceleration of $y = h(x)$ provided the product of the exponents mn is < 1. That gives us five ways to satisfy the condition of a decelerated power function between the observables.

Now suppose someone formulates a very weak theory, which specifies that there are two links to an intervening θ but does not characterize the functions further, other than stating weak constraints on their derivatives such that the input-output relation $y = h(x)$ satisfies $dy/dx > 0$, $d^2y/dx^2 < 0$, everywhere. If we are dealing with a single observational relation in a specified theoretical domain, either the logarithmic or the power theory *entails* this weak theory that speaks only about the signs of the derivatives and certain weak relations between them. So this weak theory does not "compete" for credibility with the other two; they are just stronger. They entail it, but not the other way around; and the two stronger ones do not compete with each other, because, given the experimental result, one is now falsified by the data. But considering the large class of theories over many subdomains or over different experiments even in the subdomain, we can conceive "comparisons as to credibility" on the basis of *predictive specificity*, narrowness, intolerance, riskiness, "damn strange coincidence" character of these three sorts of theories. In the set of two-link theories mediating the logarithmic prediction the theory has no competitor and, if we are willing to apply the principal of indifference, has an even chance of being correct. In the set of theories successfully mediating a predicted decelerated power function, each has one chance in five of being correct. [*Query*: Is this an illegitimate application of the principle of insufficient reason? If questioned, can we legitimate it by postulating that the scientist at least does not have a *negative* talent for zeroing in on the less probable competitor?[35]]

[35]In the social sciences, I am not sure about that as a safe assumption. As an undergraduate over half a century ago, I took a beginning course in one social science in which almost every major thesis turns out to have been erroneous!

It can easily be shown that, if the objective relative frequencies of different classes of theories are unequal, and we do not attribute to the scientist a knack of doing better than chance but at least assume that the contingency table between his distribution of choices and the objective frequencies does not have a *negative* correlation, the scientist's hit frequency will be a monotone function of the number of alternatives. But we note that the weakest class of theories which do nothing but impose very broad constraints upon the first and second derivatives of the linkage functions, so as to satisfy the general condition of monotone increasing decelerated input-output function, has no competitors so long as we confine ourselves to such a two-link situation. Here Popper is correct, that the weakest theory, the one that says the least, has the most impoverished content, and represents less of a Salmonean coincidence than either of the other kinds, also will have the highest empirical truth frequency. [*Query*: Between a weak IV theory and a strong IV theory (e.g., "monotone increasing decelerated function" and "logarithmic function" as the derived input-output relation so they cannot contradict each other and both fit the facts), is it rational for scientists to prefer the latter, as they surely would? As a *class*, the latter are less probable, but more highly corroborated in Popper's sense.]

Most theories, even in psychology, are not pure intervening variable "black box" theories, involving a construct θ (a) that is totally nondescript and (b) whose input and output relations are *simply postulated*. For a given pure IV theory, there are several accessible subtheories, and typically—in either a primitive or advanced science—several actual theories, that *derive* one or both functional IV relations from postulates concerning hypothetical entities and relations together with some postulated functional relations that are not identical with the linkage functions. These are the kinds of theoretical conjectures that have great deductive fertility, and they are of course the kind that provide most scientists with the highest level of intellectual satisfaction.[36] In the advanced physical and biological sciences these kinds of theories are the rule rather than the exception. An example is when the molar gas laws are derived from the kinetic theory of heat via mechanics plus certain assumptions about the distribution of velocities (Maxwell's derivation of the distribution of molecular velocities relied on a variant of the principle of indifference, I believe). Another example is the theory of the gene (again making plausible assumptions along principle of indifference lines in the context of what is known about the cytology of cell division) generating the statistics of population genetics. One thinks of Estes's classic derivation of the acquisition function from

[36] I do not mean in any way to denigrate pure IV theories and have contributed to such myself (e.g., MacCorquodale and Meehl's axiomatization of Tolman's expectancy theory, 1953, 1954).

the Guthrian notion that the organism samples stimulus elements which become either connected or alienated on successive occasions (Estes, 1950; Guthrie, 1935), or Biederman's derivation of tachistoscopic perception from his hypothesis of geons (Biederman, 1987). In my theory of schizophrenia, I claim to derive *all* of the major signs, symptoms, and traits of that illness by combining a postulated defect of synaptic control parameters *with other general laws already known about how the brain works*, independently of schizophrenia research (Meehl, 1962b, 1989b, 1990c, 1990d). It is needless to multiply examples, they are everywhere you look.[37] Theories of this kind often generate unexpected relations and even qualitatively novel phenomenon via a fairly complex derivation chain such that one experiences "conceptual surprise" at the result. And, of course, if such a surprising result pans out in the lab or in the clinic files, this is one of the strongest kinds of support that scientific theory receives. Nozick (being a libertarian) calls them "invisible hand" theories (1974) from the Adam Smith notion that many things happen in society as an indirect result of each individual economic atom pursuing its own selfish interests.

Although the commonest kind of novelty-generating, intellectually satisfying, invisible hand derivations are those involving the postulation of certain micro-entities and micro-events, so that *structural-compositional* theories[38] are among the most impressive in science, the introduction of unobservable entities is not a necessary condition for invisible hand effects. *Example*: I recall in one of the early operant behavior conferences which MacCorquodale and I attended at Indiana in the late 1940s, somebody reported an experiment in which one group of rats was undergoing extinction following a CRF (continuous reinforcement) schedule, and another group following a FI (fixed interval) schedule. After the operant had proceeded well along toward extinction but was still above the operant level, a "free" pellet was delivered, but carefully timed to assure that this *gratis* delivery did not follow closely upon a lever press. The experimenter invited our group to predict the effect on the cumulative record at that point. William Estes, reflecting for perhaps ten seconds, confidently stated that the effect on the animals

[37]Perhaps psychology has relatively more quasi-pure IV theories, although this is the kind of conjecture that requires a statistical sampling of literature as Faust and I advocate.

[38]Examining the various physical, biological, and social sciences, one discerns three broad classes of theories and "portions" of theories (Meehl, 1987, pp. 8-9): *Structural-compositional* (what is an entity composed of, how are its components put together? e.g., Watson-Crick gene, chemical composition of baking powder, Hebb cell assembly, Freud psychic institutions, helium nucleus); *Functional-dynamic* (what are the causal laws relating changing variables? e.g., Skinner's operant behavior theory, classical thermodynamics, Newton's mechanics, visual perception processes, Keynes' employment theory); and *Historical-developmental* (what is the sequence of states, properties, changes? e.g., big bang cosmology, Darwin's theory, Piaget's cognitive stages, Freud's libidinal stages, mammalian fetal development, Spengler's theory of history). Most scientific theories, especially in their advanced form, are integrations of these.

following the CRF schedule would be a momentary increase in rate, and after a FI schedule a momentary decrease, which was correct. We all thought that this quick response testified both to Estes's intellect and to the validity of Skinner's system!

For each pure IV theory consisting solely of postulated functional relations between input, theoretical state variable, and output there are some number k of accessible function deriving postulate sets. (Only a proper subset of these are actually proposed.) The number of accessible sets deriving the log function is $k_{\log}$. Thus a theorist who opts "randomly" for one of these sets has one chance in $k_{\log}$ of being correct. If the observed relation is a power function, and the mean number of accessible postulate sets deriving any of the 5 IV combinations is $\overline{k}_{\text{pow}}$, the theorist who employed a particular postulate set to predict that finding (within the overall requirement of only two links) has one chance in 5 $\overline{k}_{\text{pow}}$ of being right. For the weak IV theory, it is reasonable—I should think almost certain—to suppose that *some* but *not all* of the k_{weak} postulate sets capable of deriving these weak constraints on the functions are equivalent to sets from $k_{\log}$ and k_{pow} that have been "weakened" in various ways.[39] There are of course hundreds of mathematical functions in which the consequence of satisfying certain relations of the derivatives will be that the first derivative of the input-output function will be positive throughout and the second derivative negative.[40] Putting it in the other direction, *not all* of the k_{weak} sets

[39]I cannot provide a rigorous formal definition of "weakening" a postulate set and am unsure whether Popper's "consequence class that forbids less," while it has a close affinity to what we are up to here, exactly fits our situation. Obviously there are several ways to alter a theory so as to make its prediction of an observable functional form less specific. *Examples*: (a) Delete a postulate that appears in a derivation chain; (b) delete all postulates mentioning a certain entity, thus eliminating it entirely as a construct; (c) add a postulate about a variable not manipulated or directly observed that functions as a "fudge factor"; (d) strike a postulate that relates certain parameters; (e) delete a quantitative restraint on a variable or widen its allowed range; (f) split a theoretical construct into two that now appear only in different postulates with no identification of their quantitative values. The most obvious case of weakening is simply generalizing a function form occurring in a (strong) postulate, e.g., $7 + 2x^2$ "weakens" to $a + bx^2$, which weakens to "quadratic," which weakens to "increasing decelerated function," which weakens to "increasing function," which weakens to "[some] function of," a progressively weakening sequence providing 6 degrees of strength. Listing and analyzing the *qualitative* varieties of weakening is a technical task for the logician and mathematician rather than the psychologist, but a list of "levels of theory specification" such as those presented in Table 2 may be helpful as a starting orientation. I predict that a satisfactory explication of postulate set weakening will be complicated and difficult, requiring a "deep structure" mix of logical, mathematical, and bootstrapsed empirical considerations. The cliometric metatheorist must not allow logicians to impose conventional criteria in these matters—nothing is sacred except theorems of formal logic. We cannot even be sure, in advance of surveying historical episodes, whether all seeming "weakenings" affect predictive risk in the same way. For example, case (c) weakens numerical predictability of individual events *if the preweakened theory had such predictability*. But a postulate superimposing Gaussian fluctuation onto a growth function, whereby a response probability becomes the integral above a cut [e.g., Hull's (1943) behavioral oscillation fudge factor ${}_{S}O_{R}$], yields an input-output function which is equally "specific" *but rarer* and hence, on one view, a severer test (cf. p. 64 *supra*).

[40]On the IV conjecture that both input and output linkage functions are positive, the generic ("weak") condition for $y = h(x)$ to be a monotone increasing decelerated function is that

could be transformed into one of the postulate sets of the logarithmic and power function collection simply by adding one or more postulates, or by specifying certain of the microfunctions, or by constraining some parameters numerically. I cannot assert that this would *never* happen, in considering explanations of a particular experimental input-output relation. Keep in mind that we are speaking throughout of a *class* of experimental-theoretical contexts, e.g., all mammalian behavioral experiments yielding input-output functions $y = a \log x$ or $y = bx^c$, derived by two theoretically interpreted IV relations $\theta = f(x)$, $y = g(\theta)$, these functions in turn being derived from postulate sets. I repeat that I am making plausibility arguments here. It seems extremely improbable that the number of weak postulate sets capable of deriving the weak functional nondescript theory of the third kind would be equal in number and equivalent in structure, one-to-one, to the accessible postulate sets that can give rise to the power and logarithmic relationships. If this should happen to be true for a few subdomains, we are considering the truth-frequency of theories taken over a scientific domain, even if heterogeneous. I find it inconceivable that my auxiliary conjectures would fail over, say, all of psychology! Usually the accessible set k_{weak} will be quite large, and hence a given one has many more competitors than is true for the accessible postulate sets corresponding to the logarithmic and power theories. Thus, again, applying the principle of insufficient reason, unless there is some strange systematic bias in which scientists concocting or supporting stronger theories have a negative talent in choosing from the accessible postulate sets, we conclude that those hypothetical construct postulate sets that entail the weaker prediction will *as a class* have a lower truth frequency than the stronger ones. Assuming no negative bias—a perverse or stupid disposition to concoct, investigate, or prefer false theories over true ones—does *not* mean that the scientist has probability $p > ½$ of correct conjectures. It merely means that over the class of theories, whatever their truth frequencies (associated with various properties) may be, the statistical contingency table does not show a

---

$\left(\frac{d^2\theta}{dx^2}\right)\left(\frac{dy}{d\theta}\right)+\left(\frac{d^2y}{d\theta^2}\right)\left(\frac{d\theta}{dx}\right) < 0$ everywhere. Hence either of the second derivatives, or both, must be negative. If one is positive (say, the output link), we must have

$$\frac{\left|d^2\theta/dx^2\right|}{\left|d^2y/d\theta^2\right|} > \frac{\left|d\theta/dx\right|}{\left|dy/d\theta\right|}$$

i.e., the ratio of the absolute values of the second derivative of the input linkage to the second derivative of the output must exceed the ratio of the corresponding first derivatives. Obviously many sets of "common" functions can satisfy this very general condition, and, of course, the IV functions for logarithmic and power function $y = h(x)$ do so.

negative correlation between truth and preference. The scientist's individual choices may of course have truth-frequency $p \ll ½$, as it does in fact, in many (most?) sciences.

Formulating the argument rigorously, we make the assumption A that when the $k_{\log}$ accessible postulate sets deriving the input-output function $y = a \log x$ are weakened so as to derive only that $y = h(x)$ is monotone increasing decelerated and the same is done for the 5 $\bar{k}_{\text{pow}}$ sets deriving $y = ax^b$, *not all* of the latter are isomorphic with the former. Strike those in the latter that are duplicates, and let $\bar{k}$ denote the average number remaining. Then, even if the scientific community were so unimaginative that logarithmic and power functions were the only accessible monotone increasing decelerated functions (i.e., they could not conceive of any others, such as a growth function), we have for the number of weak, generic postulate sets

$$k_{\text{weak}} = 5\,\bar{k}'_{\text{pow}} + k_{\log}$$

hence, both right hand terms being $\neq 0$,

$$k_{\text{weak}} > 5\,\bar{k}'_{\text{pow}}$$

and

$$k_{\text{weak}} > k_{\log} .$$

Taking reciprocals,

$$\frac{1}{k_{\text{weak}}} < \frac{1}{5\bar{k}'_{\text{pow}}}$$

and

$$\frac{1}{k_{\text{weak}}} < \frac{1}{k_{\log}} \quad .$$

Choosing a theory from the weak set randomly, applying the indifference principle (on Assumption B = Scientific choice is not *negatively* correlated with truth frequency), a chosen weak theory is less probable than a chosen strong theory.[41]

---

[41]My colleague William Grove, with whom I co-teach a seminar on these matters, is "troubled by the use of the principle of indifference, for all the usual reasons" (personal communication, February 1992). So am I, as the text queries. One can formulate the conclusion without reliance on that principle, as an "objective" statement about truth-frequency, thus: Ranging over theories in a give domain or subdomain, the proportion of true theories—a relative truth frequency in a finite set—among the "strong" (more specific) exceeds that among the "weak" (less specific). *This formulation avoids mention of the choosing scientist*. While the theorem is *about* psychosocially accessible theories, given that minimal psychologism [a "venial sin," in the logical empiricist tradition (cf. Meehl, 1990b, pp. 35-42)], it asserts something about a relative frequency in a finite class of abstract dispositional entities, regardless of the scientist's actual choice among accessibles. We can *inform* the scientist, relying on this "objective" theorem, "If you were to randomly select more specific theories (that predict correctly), your truth-frequency would be higher than if you selected less specific ones." Were he to ignore this advice, the theorem still holds. However, adhering to our conception of metatheory as the empirical social science of scientific theorizing, we would like to conclude something stronger about empirical scientific "success" as actually conducted. Since I consider Assumption B to be a highly probable auxiliary (*over domains*, not necessarily in every subdomain or subcommunity of theorizers), and it warrants the principle of indifference algebraically, I am content until further notice to invoke it.

I cannot make this plausibility argument deductively rigorous because it is logically conceivable that when the sets of strong theories (log, power, etc.) are weakened so much as to yield only an experimental prediction of "monotone increasing decelerated," the resulting number of distinguishable postulate sets per IV combination will have been so reduced as to wash out the effect of there being several combinations. A particularly bad scenario would arise if, for some reason (difficult to imagine, but …) the experimental results greatly predominate in a function for which the number of accessible theories suffers marked "condensation" (collapse of numerous postulate sets to few) when weakened. What we must contend with here is the tension between *number of different experimental outcomes* (piling up more terms on the right side of the desired inequality) and the *condensation effects on each term*. I have been unable to simplify these sums of products to get a short, general inequality condition, and I believe it cannot be done.[42] The best I can provide is some numerical results given broad plausible conditions.

Facing the dangerous scenario, we may allow ourselves a more realistic condition on the experimental results and the scientist's creativity, but still holding each to clearly safe ("pessimistic") values. Suppose the only experimental functions that occur are five in number—logarithm, power function, simple growth function, hyperbola, reciprocal—with parameters set so as to yield a monotone increasing decelerated input/output law. Over 100 experiments in a domain, let the mean numbers of accessible postulate sets for deriving these five kinds of curves be $\overline{k}_1$, $\overline{k}_2 \ldots \overline{k}_5$. Let the condensation coefficients (that reduce the number of alternative postulate sets) be c_1, c_2 … c_5, the number of weakened sets for an outcome curve being $\overline{k}'_i = c_i \overline{k}_i$. Then the desired inequality, stating that the total number of strong theories is less than the total number of weak theories, is

$$\sum^{5} \sum^{n_i} \overline{k}_i < \sum^{5} \sum^{n_i} c_i \overline{k}_i$$

I distributed numbers of the five curve types in three ways, thus:

20	20	20	20	20
30	25	20	15	10
50	25	13	7	5

[42]Nothing in Hardy, Littlewood, and Polya's compendious treatise (1952) seems to apply, but perhaps a better mathematician than I could discern a subsumption of the present case.

The numbers of strong postulate sets were distributed in six ways, thus:

10	10	10	10	10
5	5	5	5	5
10	8	6	4	2
2	4	6	8	10
8	3	5	7	2
3	8	5	2	7

And the five condensation coefficients were distributed in nine ways, thus:

.8	.8	.8	.8	.8
.5	.5	.5	.5	.5
.2	.2	.2	.2	.2
.1	.1	.1	.1	.1
.1	.1	.1	.2	.3
.1	.1	.2	.3	.4
.1	.2	.3	.4	.5
.7	.6	.5	.4	.3
.3	.4	.5	.6	.7

Taking all possible combinations of these distributions to get 162 numerical results, we find that 90 of the 162 (= 56%) satisfy the desired inequality. The mean truth-frequency for strong theories is ${}_Tp_s = .174$, and for weak theories is ${}_Tp_w = .153$, the difference $\Delta p = .021$, and the ratio strong:weak = 1.135. This appears discouraging unless one looks at the pessimistic cases that dissatisfy the inequality, asking what configurations produce this untoward result. Since there are five experimental curve types, if the effect of weakening reduced the average number of postulate sets by a constant 80% over types, then $c = .2$ would be the break-even point, where $n_{\text{strong}} = n_{\text{weak}}$ after condensation. We find that for c-distributions where all $c_i = .8$ or all $c_i = .5$, the desired inequality holds, regardless of the curve type distribution. The same is true for variable c-distributions with $\overline{c_i} = .5$ (.3, .4, .5, .6, .7 and that sequence reversed). If the c-distribution has $\overline{c_i} = .3$ (.1, .2, .3, .4, .5 and the reversal of that sequence), the inequality fails in four configurations, *in all of which the condensation coefficients are very strongly negatively correlated with the curve type frequencies or postulate set numbers or both* ($\overline{r}$s = –1.0. –1.0, –1.0, –.92, –.92, –.92, –.50). As expected, when the condensation coefficients are $c_i \leq .2$, the adverse result usually obtains, although favorable results still can occur if the c-distribution with $\overline{c_i} = .2$ is positively correlated with the experimental curve type frequencies. For example, the c-distribution $c_i = .1, .1, .2, .3, .4$, when

associated with curve-type distribution where all $n = 20$ and postulate set numbers run 2, 4, 6, 8, 10, yields a 1.37 ratio of truth frequencies in the desired strong/weak direction. I can think of no reason for expecting these three distributions to be unfavorably correlated over a domain to such extreme degrees as the failed inequality requires, together with condensations averaging as low as $c \leq .2$. We may hope that what Einstein said applies to the class of accessible theories: "Raffiniert ist der Herrgott, aber boshaft ist er nicht."

Of course the Faust-Meehl Thesis says the proper way to study this is by empirical samplings from various research domains. For now, however, I permit myself an armchair argument about the condensation coefficient. Is it plausible to anticipate many cs $\leq .2$? [More generally and realistically, is it likely that $c \leq m^{-1}$ where m = Number of distinct experimental curve types that appear in the lab? That seems extremely far-fetched when so many different complicated functions arise in empirical research. Surely in a broad scientific domain (e.g., mammalian learning) there must occur at least 20 different input-output function types, which would yield adverse ratios only if the weakening reduced the average number of distinguishable theories per curve type by a factor of 95%, i.e., $c \leq .05$!] The pure IV theories we began with involve an input link $\theta = f(x)$ and an output link $y = g(\theta)$. These were turned into hypothetical construct, "explanatory" theories by postulating theoretical entities in a nomological network entailing the pure IV functions. Consider a conservatively small number of accessible postulate sets, say $k_{x\theta} = k_{\theta y} = 3$, for deriving each of the two IV functions in an experiment by a substantive conjecture about theoretical entities and processes involving them. Some of these entities will appear in both the $k_{x\theta}$ and the $k_{\theta y}$ sets, others will (usually) not. There are nine alternative strong theories (3^2 subtheory combinations). In order to suffer a condensation $c \leq .2$ upon weakening, there must be only a single weakened theory left for each IV link, *hence only one total theory accessible to explain the experiment*. It is hard to believe this would be the case except very rarely. Among the configurations I examined, for condensations of .8 or .5 or sets of cs averaging the latter (e.g., .3, .4, .5, .6, .7) the truth frequency *ratio* of strong to weak theories ranges from 1.79 to 4.00 with a median of 2.50, a highly satisfying result.

The reader may be uncomfortable with such nose-counting over theories in a domain, lumping experimental graphs of logs and growth functions, shrinking counts of postulate sets, etc.—the "adding lemons and oranges" worry. (As Robyn Dawes says, there is nothing wrong with that if our interest is in the class of citrus fruits.) The rationale of counting theories in the present context is identical with that of a life insurance company. The actuary classifies applicants by coarse criteria of high relevance (e.g., age, sex, weight, occupation, disease history) and assigns premium rates accordingly. What *counts* is frequency of deaths in y years, never mind that some insured

die of coronary thrombosis and others in automobile accidents. Monte Carlo is sure to remain solvent, but the house advantage differs for roulette, caged dice, and blackjack.

The above proof concerns a broad class of theories and experiments but a narrowly defined question: over-all truth frequency associated with experimental specificity. Obviously, other theory properties can countervail this property, and it would be absurd to proffer metatheoretical advice "Always bet on more specific theories." (The insurance actuary prefers young females with low blood pressure and appropriate weight—but not those in this category who drink a quart of gin daily!)

When we consider empirical truth frequency, instead of there being a negative correlation between riskiness or strength of observational content and "probability" of being true, it seems to be the other way around. Since this is in accord with the way scientists usually think, I am comfortable with this result, despite its *apparent* inconsistency with arguments by some logicians, which will be examined below.

Let me put some flesh on these bones by spinning out an imaginary state of affairs in psychology, which I hope will appear not only conceivable but plausible, to illustrate the above points. Consider some fairly broad subdomain (e.g., learning, perception, affectivity, memory), and let us imagine there are only two main groups of theorists working on explaining the facts of that domain. The one group is composed of psychologists who are knowledgeable mathematically and biologically, persons whose cognitive talents *and interests* are similar to those of the layperson's conception of "natural scientist." They favor conjectures about the way the brain works and derivations of phenomena that rely upon the mathematics of matrix algebra, probability theory, and calculus. They are busy concocting theories of this kind about various subdomains of the specified domain. Our other group are less "scientifically" inclined, perhaps more likely to belong to the APA Division of Humanistic Psychology, and, while their IQs may be equal to those of the first group, they do not have such incisive minds and they are for the most part literally uninformed about the kind of neuroscience and mathematics that the first group uses. Their explanatory concepts are commonsensical and mentalistic in character. Those who have philosophical interests would be fond of the ordinary language philosophers and the later Wittgenstein, which the scientific group would either ignore or consider inferior to the broad tradition of the logical empiricists and their off-shoots. Now the *kinds of theories* that these two groups concoct will be strikingly different, and with few exceptions the first group will tend to come up with conjectures that are relatively more quantitatively specific, partly because they will be mathematical and for that reason will have more invisible hand surprising predictions. Their theories have a stronger content, and are hence riskier and of lower absolute (prior) probability,

in the philosopher's terminology. Let us suppose, which I ask the reader to accept for purposes of argument (I myself think that it is almost certainly the case), that theoretical terms that are essentially refinements or refurbishments of mentalistic concepts do not, by and large, denote the actual "efficient causes" (Aristotle) of behavioral relationships. At best, mentalistic terms designate a special, rare subset of inner events that are reportable, and with a considerable vagueness, for reasons that have been given by Skinner (1945).[43] By far the largest part of the behavior relationships we observe are the outcome of complicated internal states and events, only a small fraction of which are accessible to introspection and have been given precise mentalistic labels in ordinary language. Hence, to put it bluntly, almost all of the *explanatory* theories of the second group of psychologists will be literally false; whereas a sizable number, perhaps even the majority, of the theories that the first group comes up with will be either literally true or have high verisimilitude. Here we have a case in which the *character* of the theories concocted—reflecting the competencies, motivations, values, and cognitive predilections of the theorizers—leads to a high positive correlation between riskiness or logical improbability as seen by the logician, and empirical truth frequency in the actual scientific world.

Does this scenario conflict with theorems of symbolic logic? It had better not, since a valid theorem of symbolic logic is just as solid as a theorem in geometry or calculus. Let me explain why there is no contradiction involved here. In ordering sets of propositions as to their logical probability (meaning, roughly, that we do not consider whatever empirical truth frequency they may have, or the factual support they may have received, but only the relations of the propositions among themselves), the philosopher relies on such truisms of symbolic logic as the following:

$$a \cdot b \supset a$$

or

$$c \cdot d \cdot e \supset c \cdot e$$

which material conditional can of course be written more strongly as a deductive entailment

$$a \cdot b \vdash a$$
$$c \cdot d \cdot e \vdash c \cdot e$$

or

$$\Box[a \cdot b \supset a]$$

[43]The verbal report of a college student experiencing a negative afterimage may be a real example of a direct causal influence from an event described in commonsensical mentalistic language, refined by words such as ⌜desaturated⌝ , but this sort of thing is the exception.

(where ⌜□⌝ is the logician's *necessary* box) since *a* is of course strictly implied (deductively entailed) by the conjunction of *a* and *b*.

Now consider two columns of theories set into one-to-one correspondence with each other, such that each of the theories on the right is obtained by striking one or more propositions from the conjunction on the left; or, putting it the other way, each theory in the left-hand column is enlarged, "strengthened," "increased in content" over its sibling in the right column, by adjoining one or more propositions to the right-hand theory. Then since there is an implication going from the left to the right, but no implication in the reverse direction, whatever may be the empirical situation—we needn't know anything at all about that—the *proportion* of theories on the left that are true must necessarily be equal or less than the *proportion* of true theories on the right. If *any* single proposition in the left column (and not found in its right column sibling) is false, the ⌜≤⌝ becomes ⌜<⌝, as will surely be true for empirical theories fitting this model. What relevance does this theorem have to the truth frequency of actual theories in empirical science? How often does it happen that the difference between two competitive theories consists simply in having added or deleted one or more postulates from a theory *T* to generate its correlated theory *T'*? Hardly ever. Thinking about psychology, I have not myself been able to come up with a single such real example, nor have any colleagues I have invited to try. One possibility, accessible but not actual, would be if one were to adjoin to the postulates comprising Skinner's theory one or two additional postulates to handle the phenomenon of latent learning, which Skinner and his disciples have never explained and almost invariably avoid discussing. As always in modifying strong theories that are doing well in order to handle a recalcitrant fact (Meehl, 1990a), one would have to be careful in formulating such a postulate not to sabotage inadvertently some previously adequate derivations in the fact domain. In the Skinner case, for instance, a postulate involving some concept of expectancy might spill over into the presently adequate derivations of the shapes of cumulative records on various schedules. I had first thought that Freud's postulation of an aggressive instinct (Freud, 1920/1955) would be a mere adjunction; but on looking again at this monograph and its successors in the 1920s, I concluded that a good deal of revision also took place in the original system, as in such concepts as the "defusion of instincts" and the mixture of aggressive and erotic components in sado-masochism. I am not claiming that there are no such examples, but apparently they are exceedingly rare, since knowledgeable persons have so much trouble thinking of any. [*Query*: Was dropping parity conservation from quantum mechanics simply striking a single postulate, leaving all the "successful derivations" intact, as I have been informed by two ex-physicist colleagues?]

There is, however, one important one kind of theorizing that does involve little more

than adjoining new postulates to a going theory, namely, successful reduction. When Watson and Crick explicated the inner nature of the gene in terms of the four bases and the backbone sugar phosphate helix, they did not have to deny or even question what was already theorized about the gene as a construct *via implicit definition* (population genetics + cytology). One would not speak of the Watson-Crick gene theory as a *competitor* of the theory of the gene as it existed in 1953. It is a stronger theory than the unreduced theory of the gene, and therefore from the logician's standpoint has a lower absolute logical probability. But the scientist is not in a position of having to choose between the two theories, he is simply asking whether the reduction is correct. [*Query*: In cases of reduction in the history of science, is the truth frequency of augmented theories in fact lower than that of the theories prior to the reduction?]

One helpful analogy is the dispute about Bernoulli's theorem (the Law of Large Numbers) as employed in the classical theory of probability. Empirical frequency theorists (e.g., Ellis, 1849, 1856; Venn, 1888) and other subsequent frequency theorists (von Mises, 1939; Reichenbach, 1938) have objected to the way that the classical theorists such as Laplace employed the Bernoulli theorem to build a bridge between the formalism of the probability calculus and the empirical frequency of properties of event-sequences in the physical world. The so-called "limit" in the frequency theories is not a limit in the usual sense of the mathematician (that is, an *analytical* convergence of a function that can be "forced" as close as you want to the limit by assigning the variable close enough to some value). Rather it is a *stochastic* convergence where we say that the probability of an empirical relative frequency diverging from the postulated true probability by more than a specified amount can be made as small as you please by taking the number of trials sufficiently large. This procedure, which works well in the application of the probability calculus to games of chance, insurance company statistics, epidemiology, genetics, and so on, lacks an adequate philosophical basis, as is generally admitted. What it comes down to is that Bernoulli's theorem, like the other theorems of the probability calculus, is simply a formal truth of combinatorics, and one does not need to be a logical positivist to wonder how a theorem of combinatorics, which would go through if completely uninterpreted in the calculus, can possibly coerce empirical frequencies in the world. It might appear that the logician's statement about the two columns described above somehow coerces the world, but it does not because all that is involved there is saying that the world will always be self-consistent. If propositions *a* and *b* and *c* on the left are true, then proposition *a* will tautologically be true. (As the early Wittgenstein said it, "[The reason the world cannot violate the laws of logic, is that] we could not *say* of an 'unlogical' world how it would look" [1922, p. 43, 3.03I].)

When the concept of absolute (logical) probability or prior probability or antecedent probability is not formulated comparatively, as in the two columns situation, I must confess I have never quite understood what it means, or how a metric is rationally assigned to strength of content. I don't mean to say that there is no such legitimate concept, but only that I do not understand it. The main thing for present purposes is that empirical theories are rarely "in competition" with others that differ from them only by the addition or subtraction of a conjunct in a conjunction of postulates. The usual situation in empirical science is not like the tautologies given above, but (except for reductions) compares theories

$$a \cdot b \cdot c \text{ vs } a \cdot d \cdot f$$

or

$$a \cdot b \cdot c \cdot d \text{ vs } f \cdot g \cdot h\,,$$

and between these competitors no logical relation obtains. The logician's unexceptionable claim about the inequality holding between sets of theories differing in such a way would not contradict an empirical finding that theories stronger in the sense of making more detailed or precise factual observational predictions have, in the history of science, a higher truth frequency than weaker or looser theories. [*Query*: Can cognitive psychology contribute to our understanding of types of scientific minds in relationship to the kinds of theories they tend to devise?]

WHO SHOULD DO CLIOMETRIC METATHEORY AND HOW SHOULD THEY BEGIN?

It would be presumptuous of me, having expertise in neither history of science nor the quantitative study of scientific communication, to lay out a detailed blueprint for conducting cliometric metatheory. Nor do I wish to specify which of the scholarly professions should take the lead. In starting an enterprise so novel and so deviant, with almost no empirical data to go on, only a sketch of plausible initiating steps is possible *or necessary*. A few exemplary studies by a small number of workers should suffice to get us going, or to show that it's a poor idea. But I warn against drawing a pessimistic conclusion prematurely; I would set high qualitative standards on preliminary studies if they are to be used as litmus tests of the approach. Naturally the quality and amount of exploratory research one requires before concluding adversely will depend on how persuasive the arguments and evidence presented in this paper seem.[44] I offer—not as a

[44]I am highly confident ("personalistic subjective probability"!) that the core thesis is essentially correct, and I would not throw in the sponge readily; but of course I cannot expect others to feel as I do. My co-author on a related paper, David Faust, has lectured on the topic about a dozen times and regularly finds the reactions of philosophers, historians, and sociologists to be either misunderstanding or comprehension combined with hostility (personal communication, September 1991). Interestingly, the only professional group that has reacted with sympathetic understanding are seasoned CEOs! Their refreshing reason: "I have learned the painful way, losing money, to trust the figures more than my 'expert' judgment."

recommendation but as social prediction—the opinion that psychologists are most likely to be early explorers of this new technology, historians of science next, and philosophers least likely. Psychologists' familiarity with the clinical-statistical controversy and with the research on cognitive processes, attribution, judgment, etc., and their technical knowledge of—*and comfortable feelings with*— statistics, psychometrics, rating scales, and quantitative methods generally (especially as applied to "fuzzy problems" and subjective human appraisals), tend to make us the likely leaders in developing cliometric metatheory. That is why the present paper appears here, rather than in a philosophers' or historians' journal.

Perhaps the ideal cliometric metatheorist of the future will be a polymath, competent in statistics, psychometrics, logic, philosophy, and historiography. Arguably this interdisciplinary expert should also be knowledgeable about the particular branch of science chosen for metatheoretical study. (Today, doctoral candidates in history or philosophy of science are usually expected to acquire at least M.A. level competence in the area of science they intend to research.) However, it is unrealistic to expect people to invest such extra effort as prerequisite for "exploring" a possibly unrewarding field, especially one that most old-guard experts find uninteresting or even threatening. My suggestion is that a psychologist could contribute worthwhile exploratory work, without devoting years to acquiring such interdisciplinary expertise, either by examining theories in the "mature science" subdomains of psychology, or by teaming up with a scientist in some other science to be studied and confining the data base to scientific journals, treatises, and standard textbooks. (We assume the traditional, nonstatistical historians of science will continue to go about their case study business, and we will rely on their results as fruitful sources of metatheoretical conjectures, to be tested cliometrically.)

Suggestions as to quantitative indexes, time spans, etc., have been presented in the text supra, by Faust (1984), by Meehl (1990a, 1990b), and by Faust and Meehl (1992). It is obviously desirable to diversify over sciences. I think we should avoid the "Big Theories" (produced by scientific revolutions) such as Kepler, Newton, Virchow, Darwin, Koch, Freud, Skinner, Einstein, and quantum mechanics, and concentrate instead on mini-theories (e.g., the benzene ring, how chlorophyll works, continental drift, diabetic physiology, transmission of tuberculosis, genetic linkage and chromosome maps, classical conditioning, capillary attraction, solid state physics, fluid mechanics). I believe one of the worst curses on philosophy of science has been excessive concentration on "spooky" questions arising in quantum mechanics; and it seems generally agreed today that the tendency to treat theoretical physics as the exemplar of all science is a bad mistake if we want to metatheorize about the biological and social sciences.

A random selection of middle level theories from a specified domain, say biochemistry, could be done by starting with standard treatises and textbooks, with fairly quick and easy elimination of candidates that survived for too short a time to accumulate many experiments. At this stage I would not obsess perfectionistically about complete journal coverage, but would rely on the small number of mainline, widely read journals as data base. I conjecture that a few minor excursions into the vast peripheral literature would suffice to set safe bounds on the quantitative impact of including all of it in the population to be sampled from. Excessive concern about completeness will, I believe, usually stem from the worrier's forgetting that the core idea of the cliometric program is *actuarial*, that we are not drawing our conclusions from the internal relations discernible within a case study but rather from the statistical trends exhibited by an aggregate of cases whose attributes can be "objectively" tallied with minimal inference about connections among these "innards."

If a dozen theories ranging over geology, physiology, electrochemistry, epidemiology, genetics, psychology, etc., showed a negligible correlation between the time graph of my *C** index and long-term survival in the textbooks, I would abandon the index. The main thing is to keep our eye on the ball, to remember the phase we are in and what questions we are asking *appropriate for that phase of a new science*. For me, the first big question would be whether a few (crude but plausible) indexes of theory performance, such as *C**, display orderly changes over time, and correlate with theories' long-term fate and with each other as a bootstrap-coherency indication of verisimilitude. If these results are encouraging over-all, I would next be interested in comparing the actuarial method's accuracy with the impressionistic judgments that were expressed by contemporaneous scientists, and also with ratings experimentally obtained by today's scientists and philosophers, looking backward, as to the *then available support*. While I am aware that this line of investigation may reflect my longstanding interest in the clinical/statistical comparison, *I insist that Faust* (1984) *was right in seeing that issue*—extensively researched because of its practical importance, challenge to practitioners, and theoretical interest—*as being merely a striking exemplar of a highly general problem concerning human cognitive function*.

If a few properly conducted studies tend to corroborate the Faust-Meehl Strong Actuarial Thesis, one would hope that the tremendous utility of the approach in scientific, technological, and financial terms (Faust & Meehl, 1992) would lead fund-granting agencies to invest in large-scale projects. I do not think it exaggerated to claim that a significant improvement in scientific theorizing and theory-appraisal warrants as heavy expenditure as mapping the human genome or completing the catalog of stars.

RESUMÉ

1. "Scientific method" is a heterogeneous collection of rules (very few, strictly speaking), principles, guidelines, heuristics, preferences, policies, standards, criteria, followed in varying degrees by scientists, sometimes successfully, sometimes not. It is hard to state a single rule or principle that has never been violated with impunity, or even making for success on occasion.

2. No scientist, logician, or philosopher has derived all, or most, of these principles from a single postulate about the world (e.g., "Nature is lawful") or a statement of an epistemic aim (e.g., "Discover the truth"), and there is no reason for supposing such a derivation is possible. Simple inspection of available lists of principles shows that they are not interderivable, and some pairs are in opposition.

3. Whether a principle is sound—tends to facilitate approaching an epistemic aim—is an empirical question, answerable only by history of science.

4. But since a sound principle only *tends* to success, the correct empirical formulation of this fact is unavoidably actuarial. Hence an adequate formulation of the principle must specify its quantitative *validity* (e.g., correlation, hit/miss ratio, weight).

5. Episodes in history of science may, and typically do, fall under multiple principles whose individual applications, if treated qualitatively (dichotomously—subsumable? applies or not?) would yield contradictory advice.

6. Hence the scientist's decision or action involves cognitive processes of combining quantitative information whose components are of varying weight and usually tend to countervail each other.

7. The cognitive problem presented is similar—in most respects identical—to that faced by a clinician diagnosing and prognosing a patient.

8. Contrary to the conventional wisdom, learned and experienced clinicians perform this kind of task rather inefficiently, and their long-run statistical accuracy is almost never better than that of a simple nonoptimizing regression equation or actuarial table, often worse. There are now some 175 research studies, involving a wide variety of prediction domains, revealing very few exceptions to this generalization.

9. Whether an optimal (or satisficing) algorithm for combining scientific rules and principles is linear, nonlinear, interactive, step-functional, compensatory, successive hurdles, mixed model, or whatever are empirical questions. It should not be assumed that more variables or greater complexity of their combination function somehow gives the human impressionistic judge a cognitive edge over a formal ("mechanical," algorithmic) procedure based on actuarial experience. If anything, more variables and more complex relations makes subjective judgment worse.

10. Recognition that metatheory is an empirical discipline (literally a *social science*, as Sneed says) does not preclude interest in "rational" internal considerations taken

from formal logic, probability theory, or armchair epistemology. All social sciences permit putting "why" questions whose answers may—or may not—involve the simple fact that humans intend, think, know, infer, adapt means to ends, which stones and daffodils do not. The categories "valid" versus "fallacious," "appropriate" versus "counterproductive," do not disappear from our toolkit when we see that metatheory is a social science. An economist studying business failures allows himself to identify a firm's irrational decisions; a political scientist contrasts the potency of rational and irrational appeals to voters; a psychotherapist calls attention to the patient's defective reality-testing. We want our empirical metatheory to explain why science "works better" than superstition, and that cannot be done if we impose a needless proscription against invoking concepts of rationality, probability, and the like.

11. While rational considerations, probability theory, plausibility arguments, the conventional wisdom of working scientists, consensus epistemology, and cognitive sciences are proper sources of metatheoretical conjectures (how to sample? what to count? what statistical analysis to apply?), whether the Strong Actuarial Thesis is sound or not is an empirical question *and must itself be appraised by appropriate actuarial methods*.

ACKNOWLEDGEMENTS

It is difficult to trace my intellectual indebtedness in this paper, especially when I lack documents and must rely on memory. I believe my first exposure to the notion of "statisticizing" properties of theories in relation to their long-run success was in Reichenbach's *Experience and Prediction* (1938) which I read and discussed with George H. Collier a year or two after it appeared. Defending his identity conception of probability against the disparity conception of Carnap and others, Reichenbach recognizes that we do not easily understand a theory's probability as the limit of a relative frequency (cf. Nagel, 1939, pp. 65-66). But he argues that *in principle* it can be so conceived, the truth-frequency of certain classes of theories being the desired probability number *p*. The passage (p. 399) is brief and cryptic, and one cannot even tell whether the "theory properties" defining subclasses of the reference class are substantive, formal, factual, or a combination of these. The properties are to be correlated with long-term "success" frequency. Reichenbach's student Wesley Salmon never heard Reichenbach spell it out further (personal communication, August 15, 1989). In the early days of the Minnesota Center for Philosophy of Science (1953-1956), the appearance of my book on clinical prediction led us (Feigl, Meehl, Scriven, Sellars) briefly to consider the idea, but we were still so "positivistic" that it was rejected as smacking of psychologism. A few years later, Grover Maxwell's formulation of objective Bayesian-ism revived my interest. I note that in a lecture given at the Center's 1980 conference

on theory testing (published in Meehl, 1983) I make the actuarial argument explicitly (pp. 371-372) in criticizing Feyerabend and Popper for their reliance on the nonactuarial case study method as proof of metatheoretical generalizations. The strongest influence impelling me to revive the idea of actuarial metatheory comes from my friend and former colleague, David Faust, whose important book (insufficiently noticed by philosophers and psychologists) I read in manuscript (see Meehl, 1984, in Faust, 1984), and with whom I have had many profitable exchanges in correspondence and conversation. He single-handedly "cured" me of positivist remnants of psychologism-phobia, although it is testimony to the grip of that fear that, four years after my lecture's criticism of anecdotal, nonactuarial use of case study instances to prove metatheoretical generalizations, in my generally enthusiastic foreword to his book I was still somewhat ambivalent. I trace this inconsistency partly to a difference in our evaluation of Kuhn and partly to unclarity about the relation between cliometric appraisal of (ordinary) scientific theories and that of metatheories, discussed in this article. Scholarly justice requires emphasis that the pro-actuarial statements in Reichenbach, in my 1980 conference lecture (Meehl, 1983), and those made by other writers cited in Footnote 4, while explicit, consisted of only a sentence or short paragraph, whereas Faust devoted an entire book (1984) to developing the ramified position, integrating evidence and theoretical argument from several knowledge domains.

It is obvious—but to preclude any possible misunderstanding I say it explicitly—that no priority is being claimed for the general idea of statisticizing aspects of the history of science. Historians (e.g., Price, 1963) and sociologists (e.g., Merton, 1973) of science have been doing that for years, as have students of communication and documentation. The best quantitative historical research on personal creativity in science, music, literature, and politics is by psychologist Dean Keith Simonton (1990). The international periodical *Scientometrics*, self-described broadly as concerned with "all quantitative aspects of the science of science," focuses on statistical studies and mathematical models of scientific productivity, influence, and communication (e.g., an index of the journal impact factor, comparisons of national research production, paper outputs by subject matter domain, institutional networks, cluster identification of scientific "schools," measures of individual scientists' influence, judgmental ratings of articles' significance, research and development support). I have not seen papers therein quantifying properties or performance of theories or statistically testing metatheoretical principles.

REFERENCES

Allport, G. W., & Odbert, H. S. (1936) Trait-names: a psycho-lexical study. *Psychological Monographs*, 47, 1-171.

Andreski, S. (1972) *Social sciences as sorcery*. London: Deutsch.

Annual Report of the [British] *Medical Research Council* (1953-1954). Blindness in premature infants. London: Her Majesty's Stationary Office. Pp. 15-18.

Arkes, H. R. (1991) Costs and benefits of judgment errors: implications for debiasing. *Psychological Bulletin*, 110, 486-498.

Arkes, H. R., & Hammond, K. R. (1986) *Judgment and decision making: an interdisciplinary reader*. New York: Cambridge Univer. Press.

Arkes, H. R., Wortmann, R. L., Saville, P. D., & Harkness, A. R. (1981) Hindsight bias among physicians weighing the likelihood of diagnoses. *Journal of Applied Psychology*, 66, 252-254.

Atkinson, D. R., Furlong, M. J., & Wampold, B. E. (1982) Statistical significance, reviewer evaluations, and the scientific process: is there a (statistically) significant relationship? *Journal of Counseling Psychology*, 29, 189–194.

Aydelotte, W. O. (1971) *Quantification in history*. Reading, MA: Addison-Wesley.

Aydelotte, W. O., Bogue, A. G., & Fogel, R. W. (Eds.) (1972) *The dimensions of quantitative research in history*. Princeton, NJ: Princeton Univer. Press.

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

Barnes, B. (1974) *Scientific knowledge and sociological theory*. London: Routledge & Kegan Paul.

Barnes, B. (1977) *Interests and the growth of knowledge*. London: Routledge & Kegan Paul.

Barzun, J. (1974) *Clio and the doctors: psycho-history, quanto-history, and history*. Chicago, IL: Univer. of Chicago Press.

Benson, L. (1972) *Toward the scientific study of history*. Philadelphia, PA: Lippincott.

Biederman, I. (1987) Recognition-by-components: a theory of human image understanding. *Psychological Review*, 94, 115-147.

Bloor, D. (1976) *Knowledge and social imagery*. London: Routledge & Kegan Paul.

Blumberg, A. E., & Feigl, H. (1931) Logical positivism. *Journal of Philosophy*, 28, 281-296.

Bogue, A. G., (1983) *Clio and the bitch goddess: quantification in American political history*. Beverly Hills, CA: Sage.

Brush, S. G. (1974) Should the history of science be rated X? *Science*, 183, 1164-1172.

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

Burt, C. (1950) The influence of differential weighting. *British Journal of Psychology*, Statistical Section, 3, 105-123.

Camerer, C. (1981) General conditions for the success of bootstrapping models. *Organizational Behavior & Human Performance*, 27, 411-422.

Carnap, R. (1945) The two concepts of probability. *Philosophy and Phenomenological Research*, 5, 513–532. [Reprinted in H. Feigl & W. Sellars (Eds.), *Readings in philosophical analysis*. New York: Appleton-Century-Crofts, 1949. Pp. 330-348.

Reprinted in H. Feigl & M. Broadbeck (Eds.), *Readings in the philosophy of science*. New York: Appleton-Century-Crofts, 1953. Pp. 438–455.]

Carnap, R. (1949) Truth and confirmation. In H. Feigl & W. Sellars (Eds.), *Readings in philosophical analysis*. New York: Appleton-Century-Crofts, 1949. Pp. 119-127.

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

Christie, G. (1968) The model of principles. *Duke Law Journal*, 649–669.

Cofer, C. N., & Appley, M. H. (1964) *Motivation: theory and research*. New York: Wiley.

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

Comte, A. (1830-42/1974) [S. Andreski (Ed.) & M. Clarke (Trans.)] *The essential Comte: selected from Cours de philosophie positive by Auguste Comte first published in Paris 1830-42*. New York: Barnes & Noble.

Comte, A. (1830-54/1983) (G. Lenzer, Ed.) *Auguste Comte and positivism: the essential writings (1830-1854)*. Chicago, IL: Univer. of Chicago Press.

Conrad, A. H., & Meyer, J. R. (1964) *The economics of slavery*. Chicago, IL: Aldine.
Copi, I. M. (1961) *Introduction to logic*. (2nd ed.) New York: Macmillan.
Cronbach, L. J., (1951) Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297-334.
Cronbach, L. J., & Meehl, P. E. (1955) Construct validity in psychological tests. *Psychological Bulletin*, 52, 281-302. [Reprinted in P. E. Meehl, *Psychodiagnosis: selected papers*. Minneapolis, MN: Univer. of Minnesota Press, 1973. Pp. 3-31.]
Dauer, F. W. (1989) *Critical thinking: an introduction to reasoning*. New York: Oxford Univer. Press.
Dawes, R. M. (1970) *An inequality concerning correlation of composites vs. composites of correlations*. Oregon Research Institute, Methodological Note, 1(No. 1).
Dawes, R. M. (1988) *Rational choice in an uncertain world*. Chicago, IL: Harcourt Brace Jovanovich.
Dawes, R. M., Faust, D., & Meehl, P. E. (1989) Clinical versus actuarial judgment. *Science*, 243, 1668-1674.
Diamond, A. M., Jr. (1980) Age and the acceptance of cliometrics. *Journal of Economic History*, 40, 838-841.
Diamond, A. M., Jr. (1988) The polywater episode and the appraisal of theories. In A. Donovan, L. Laudan, & R. Laudan (Eds.), *Scrutinizing science: empirical studies of scientific change*. Boston, MA: Kluwer Academic Publishers. Pp. 181-198.
Donovan, A., Laudan, L., & Laudan, R. (1988) *Scrutinizing science: empirical studies of scientific change*. Boston, MA: Kluwer Academic Publishers.
Du Bois-Reymond, E. (1880/1912) Die sieben Welträtsel. In Estelle Du Bois-Reymond (Ed.), *Reden von Emil Du Bois-Reymond*. Vol. 2. Leipzig: Veit. Pp. 65-98. (Original work published 1880)
Dworkin, R. M. (1967) *The model of rules*. University of Chicago Law Review, 35, 14-46.
Earman, J., & Glymour, C. (1980) Relativity and eclipses: the British eclipse expeditions of 1919 and their predecessors. *Historical Studies in the Physical Sciences*, 11, 49-85.
Ellis, R. L. (1849) On the foundations of the theory of probabilities. *Transactions of the Cambridge Philosophical Society*, 8, 1-6.
Ellis, R. L. (1856) Remarks on the fundamental principle of the theory of probabilities. *Transactions of the Cambridge Philosophical Society*, 9, 605-607.
Erickson, C. (1975) Quantitative history. *American Historical Review*, 80, 351-365.
Estes, W. K. (1950) Toward a statistical theory of learning. *Psychological Review*, 57, 94-107.
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.
Feigl, H. (1929) Meaning and validity of physical theories. Translated and reprinted in R. S. Cohen (Ed.), Herbert Feigl. *Inquiries and provocations: selected writings 1929-1974*. Boston, MA: D. Reidel, 1981. Pp. 116-144. [Original publication in German]
Feigl, H. (1950) Existential hypotheses. 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.
Festinger, L. (1957) *A theory of cognitive dissonance*. Evanston, IL: Row, Peterson.
Feyerabend, P. (1965) Problems of empiricism. In R. G. Colodny (Ed.), *Beyond the edge of certainty*. Englewood Cliffs, NJ: Prentice-Hall. Pp. 145-260.
Feyerabend, P. (1970) 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, 1988.]
Feyerabend, P. (1971) Problems of empiricism, Part II. In R. G. Colodny (Ed.), *The nature and function of scientific theories*. Pittsburgh, PA: Univer. of Pittsburgh Press. Pp. 275-353.

Feyerabend, P. (1987) *Farewell to reason*. London: Verso.
Feynman, R. (1986) *Surely you're joking, Mr. Feynman!* New York: Bantam.
Fisher, R. A. (1970) *Statistical methods for research workers*. (14th ed.) Edinburgh: Oliver & Boyd.
Fisher, R. A. (1971) *The design of experiments*. (9th ed.) New York: Hafner.
Fitch, N. (1984) Statistical fantasies and historical facts: history in crisis and its methodological implications. *Historical Methods*, 17, 239-254.
Flanigan, W. H. (1984) The conduct of inquiry in social science history. *Social Science History*, 8, 323-339.
Floud, R. (1973) *An introduction to quantitative methods for historians*. London: Methuen.
Floud, R. (1984) Quantitative history and people's history: two methods in conflict? *Social Science History*, 8, 151-168.
Fogel, R. W. (1975) The limits of quantitative methods in history. *American Historical Review*, 80, 329-350.
Fogel, R. W., & Elton, G. R. (1983) *Which road to the past? Two views of history*. New Haven, CT: Yale Univer. Press.
Fogel, R. W., & Engerman, S. L. (1974) *Time on the cross*. Boston, MA: Little Brown.
Frank, P. G. (1954) The variety of reasons for the acceptance of scientific theories. *Scientific Monthly*, 79, 139-145.
Freud, S. (1896/1962) The aetiology of hysteria. In J. Strachey (Ed. & Trans.), *Standard edition of the complete psychological works of Sigmund Freud*. Vol. 3. London: Hogarth. Pp. 191-221 (Original work published 1896)
Freud, S. (1900/1953) The interpretation of dreams. In J. Strachey (Ed. & Trans.), *Standard edition of the complete psychological works of Sigmund Freud*. Vols. 4 and 5. London: Hogarth. (Original work published 1900)
Freud, S. (1920/1955) Beyond the pleasure principle. In J. Strachey (Ed. & Trans.), *Standard edition of the complete psychological works of Sigmund Freud*. Vol. 18. London: Hogarth. Pp. 7–64. (Original work published 1920)
Giere, R. N. (1969) Bayesian statistics and biased procedures. *Synthese*, 20, 371-387.
Giere, R. N. (1983) Testing theoretical hypotheses. In J. Earman (Ed.), *Minnesota studies in the philosophy of science*. Vol. X. *Testing scientific theories*. Minneapolis, MN: Univer. of Minnesota Press. Pp. 269-298.
Giere, R. N. (1984) *Understanding scientific reasoning*. New York: Holt, Rinehart, and Winston.
Giere, R. N. (1988) *Explaining science: a cognitive approach*. Chicago: Univer. of Chicago Press.
Glass, G. V. (1976) Primary, secondary, and meta-analysis of research. Presidential address presented at Annual Meeting of the American Educational Research Association, San Francisco.
Glass, G. V., McGaw, B., & Smith, M. L. (1981) *Meta–analysis in social research*. Beverly Hills, CA: Sage.
Glymour, C. (1980) *Theory and evidence*. Princeton, NJ: Princeton Univer. Press.
Goldberg, L. R. (1970) Man versus model of man: a rationale, plus some evidence, for a method of improving on clinical inferences. *Psychological Bulletin*, 73, 422-432.
Goldstick, D., & O'Neill, B. (1988) "Truer." *Philosophy of Science*, 55, 583–597.
Goodstein, L. D., & Brazis, K. L. (1970) Psychology of the scientist: XXX. Credibility of psychologists: an empirical study. *Psychological Reports*, 27, 835-838.
Greenwald, A. G. (1975) Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, 82, 1-20.
Greenwald, A. G., & Pratkanis, A. R. (1988) On the use of "theory" and the usefulness of theory. *Psychological Review*, 95, 575-579.
Greenwald, A. G., Pratkanis, A. R., Leippe, M. R., & Baumgardner, M. H. (1986) Under what conditions does theory obstruct research progress? *Psychological Review*, 93, 216-229.
Grünbaum, A. (1960) The Duhemian argument. *Philosophy of Science*, 27, 75-87.
Guilford, J. P. (1954) *Psychometric methods*. New York: McGraw-Hill.
Guthrie, E. R. (1935) *The psychology of learning*. New York: Harper.

Hardy, G. H., Littlewood, J. E., & Polya, G. (1952) *Inequalities*. (2nd ed.) Cambridge, Eng.: University Press.

Harris, R. J. (1975) *A primer of multivariate statistics*. New York: Academic Press.

Hawkins, S. A., Hastie, R. (1990) Hindsight: biased judgments of past events after the outcomes are known. *Psychological Bulletin*, 107, 311-327.

Hawthorne, J. (1989) Giving up judgment empiricism: the Bayesian epistemology of Bertrand Russell and Grover Maxwell. In C. W. Savage & C. A. Anderson (Eds.), *Minnesota studies in the philosophy of science*. Vol. XII. *Rereading Russell: essays in Bertrand Russell's metaphysics and epistemology*. Minneapolis, MN: Univer. of Minnesota Press. Pp. 234-248.

Hays, S. P. (1984) Scientific versus traditional history. *Historical Methods*, 17, 75-78.

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

Hilpinen, R. (1976) Approximate truth and truthlikeness. In M. Przélecki, K. Szaniawski, & R. Wójcicki (Eds.), *Formal methods in the methodology of empirical sciences*. Dordrecht: Ossolineum, Wroclaw, and D. Reidel. Pp. 19-42.

Himmelfarb, G. (1987) *The new history and the old*. Cambridge, MA: Belknap Press of Harvard Univer. Press.

Hobsbawm, E. J. (1980) The revival of narrative: some comments. *Past and present*, 86, 3-8.

Hogarth, R. M. (1987) *Judgement and choice: the psychology of decision*. New York: Wiley.

Hollingshead, A. B., & Redlich, F. C. (1958) *Social class and mental illness*. New York: Wiley.

Howson, C., & Franklin, A. (1991) Maher, Mendeleev and Bayesianism. *Philosophy of Science*, 58, 574-585.

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

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982) *Meta-analysis: cumulating research findings across studies*. Vol. 4. *Studying organizations: innovations in methodology*. Beverly Hills, CA: Sage.

Jarausch, K. H. (1984) The international dimension of quantitative history. *Social Science History*, 8, 123-132.

Judt, T. (1979) A clown in regal purple: social history and the historians. *History Workshop*, issue 7, 66-94.

Kahneman, D., Slovic, P., & Tversky, A. (Eds.) (1982) *Judgments under uncertainty: heuristics and biases*. Cambridge, Eng.: Cambridge Univer. Press.

Kapp, R. O. (1940) *Science vs. materialism*. London: Methuen.

Kapp, R. O. (1951) *Mind, life, and body*. London: Constable.

Keller, E. F. (1983) *A feeling for the organism: the life and work of Barbara McClintock*. San Francisco, CA: Freeman.

Kelly, K. T., & Glymour, C. (1989) Convergence to the truth and nothing but the truth. *Philosophy of Science*, 56, 185-220.

Kenny, A. (1983) *Thomas More*. New York: Oxford Univer. Press.

Kern, L. H., Mirels, H. L., & Hinshaw, V. G. (1983) Scientists' understanding of propositional logic: an experimental investigation. *Social Studies of Science*, 13, 131-146.

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

Kleinmuntz, B. (1990) Why we still use our heads instead of formulas. *Psychological Bulletin*, 107, 296-310.

Kocka, J. (1984) Theories and quantification in history. *Social Science History*, 8, 169-178.

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.

Kousser, J. M. (1984) The revivalism of narrative: a response to recent criticisms of quantitative history. *Social Science History*, 8, 133-149.

Kuhn, T. S. (1970) *The structure of scientific revolutions*. (2nd ed.) *International Encyclopedia of Unified Science*, 2(2). Chicago, IL: Univer. of Chicago Press.

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.]

Lakatos, I. (1970) Falsification and the methodology of scientific research programmes. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the growth of knowledge*. Cambridge, Eng.: Cambridge Univer. Press. Pp. 91-195. [Reprinted in J. Worrall & G. Currie (Eds.), *Imre Lakatos: philosophical papers*. Vol. I. *The methodology of scientific research programmes*. New York: Cambridge Univer. Press, 1978. Pp. 8-101.]

Latour, B., & Woolgar, S. (1979) *Laboratory life: the social construction of scientific facts*. Beverly Hills, CA: Sage.

Laudan, L. (1977) *Progress and its problems: toward a theory of scientific growth*. Berkeley, CA: Univer. of California Press.

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

Laudan, L. (1990) Normative naturalism. *Philosophy of Science*, 57, 44-59.

Laudan, L., Donovan, A., Laudan, R., Barker, P., Brown, H., Leplin, J., Thagard, P., &
Wykstra, S. (1986) Scientific change: philosophical models and historical research. *Synthese*, 69, 141-223.

Lehman, D. R., Lempert, R. O., & Nisbett, R. E. (1988) The effects of graduate training on reasoning: formal discipline and thinking about everyday-life events. *American Psychologist*, 43, 431-442.

Levi, I. (1967) *Gambling with truth*. New York: Knopf.

Lord, C. G., Ross, L., & Lepper, M. R. (1979) Biased assimilation and attitude polarization: the effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, 37, 2098–2109.

Lorwin, V. R., & Price, J. M. (Eds.) (1972) *The dimensions of the past: materials, problems, and opportunities for quantitative work in history*. New Haven, CT: Yale Univer. Press.

Lykken, D. T. (1991) What's wrong with psychology anyway? In D. Cicchetti & W. M. Grove (Eds.), *Thinking clearly about psychology*. Vol. 1. *Matters of public interest*. Minneapolis, MN: Univer. of Minnesota Press. Pp. 3-39.

MacCorquodale, K., & Meehl, P. E. (1948) On a distinction between hypothetical constructs and intervening variables. *Psychological Review*, 55, 95–107. [Reprinted in Meehl, *Selected philosophical and methodological papers* (C. A. Anderson & K. Gunderson, Eds.). Minneapolis, MN: Univer. of Minnesota Press, 1991. Pp. 249-264.]

MacCorquodale, K., & Meehl, P. E. (1953) Preliminary suggestions as to a formalization of expectancy theory. *Psychological Review*, 60, 55-63.

MacCorquodale, K., & Meehl, P. E. (1954) E. C. Tolman. In W. K. Estes, S. Koch, K. MacCorquodale, P. E. Meehl, C. G. Mueller, W. N. Schoenfeld, & W. S. Verplanck, *Modern learning theory*. New York: Appleton-Century-Crofts. Pp. 177-266.

Maher, P. (1988) Prediction, accommodation, and the logic of discovery. In A. Fine & J. Leplin (Eds.), *PSA 1988*. Vol. 1. East Lansing, MI: Philosophy of Science Association. Pp. 273-285.

Maher, P. (1990) How prediction enhances confirmation. In M. Dunn & A. Gupta (Eds.), *Truth or consequences*. Dordrecht: Kluwer. Pp. 327-343.

Mahoney, M. J. (1976) *Scientist as subject: the psychological imperative*. Cambridge, MA: Bollinger.

Mahoney, M. J. (1977) Publication prejudices: an experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research*, 1, 161-175.

Mahoney, M. J. (1979) Psychology of the scientist: an evaluative review. *Social Studies of Science*, 9, 349-375.

Mahoney, M. J., & Kimper, T. P. (1976) From ethics to logic: a survey of scientists. In M. J. Mahoney, *Scientist as subject: the psychological imperative*. Cambridge, MA: Bollinger. Pp. 187-193.

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

Margoshes, A., & Litt, S. (1965) Psychology of the scientist: XII. Neglect of revolutionary ideas in psychology. *Psychological Reports*, 16, 621-624.

Maxwell, G. (1974) The later Russell: philosophical revolutionary. In G. Nakhnikian (Ed.), *Russell's philosophy*. London: Duckworth. Pp. 169-82.

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

McCormack, R. L. (1956) A criticism of studies comparing item-weighting methods. *Journal of Applied Psychology*, 40, 343-344.

McMurray, G. A. (1955) Congenital insensitivity to pain and its implications for motivational theory. *Canadian Journal of Psychology*, 9, 121-131.

Meehl, P. E. (1954) *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. (1955) Psychotherapy. *Annual Review of Psychology*, 6, 357-378.

Meehl, P. E. (1962a) Psychopathology and purpose. In P. Hoch & J. Zubin (Eds.), *The future of psychiatry*. New York: Grune and Stratton. Pp. 61-69. [Reprinted in Meehl, *Selected philosophical and methodological papers* (C. A. Anderson & K. Gunderson, Eds.). Minneapolis, MN: Univer. of Minnesota Press, 1991. Pp. 265-271.]

Meehl, P. E. (1962b) Schizotaxia, schizotypy, schizophrenia. *American Psychologist*, 17, 827-838. [Reprinted in Meehl, *Psychodiagnosis: selected papers*. Minneapolis, MN: Univer. of Minnesota Press, 1973. Pp. 135-155).]

Meehl, P. E. (1970) Psychological determinism and human rationality: a psychologist's reactions to Professor Karl Popper's 'Of clouds and clocks'. 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. 310-372. [Reprinted in C. A. Anderson & K. Gunderson (Eds.), *Selected philosophical and methodological papers* [of P. E. Meehl]. Minneapolis, MN: Univer. of Minnesota Press, 1991. Pp. 43-96.]

Meehl, P. E. (1971) Law and the fireside inductions: some reflections of a clinical psychologist. *Journal of Social Issues*, 27, 65-100. [Reprinted, abridged and with Postscript, *Behavioral Sciences and the Law*, 1989, 7, 521-550.]

Meehl, P. E. (1973) *Psychodiagnosis: selected papers*. Minneapolis, MN: Univer. of Minnesota Press.

Meehl, P. E. (1977) The selfish voter paradox and the thrown-away vote argument. *American Political Science Review*, 71, 11-30.

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. [Reprinted in *Meehl, Selected philosophical and methodological papers* (C. A. Anderson & K. Gunderson, Eds.). Minneapolis, MN: Univer. of Minnesota Press, 1991. Pp. 1-43.]

Meehl, P. E. (1979) A funny thing happened to us on the way to the latent entities. *Journal of Personality Assessment*, 43, 563-581.

Meehl, P. E. (1983) Subjectivity in psychoanalytic inference: The nagging persistence of Wilhelm Fliess's Achensee question. In J. Earman (Ed.), *Minnesota studies in the philosophy of science*. Vol. X. *Testing scientific theories*. Minneapolis, MN: Univer. of Minnesota Press. Pp. 349-411. [Reprinted in Meehl, *Selected philosophical and methodological papers* (C. A. Anderson & K. Gunderson, Eds.). Minneapolis, MN: Univer. of Minnesota Press, 1991. Pp. 284-337.]

Meehl, P. E. (1984) Foreword. In D. Faust, *The limits of scientific reasoning*. Minneapolis. MN: Univer. of Minnesota Press.

Meehl, P. E. (1986) Causes and effects of my disturbing little book. *Journal of Personality Assessment*, 50, 370-375.

Meehl, P. E. (1987) Theory and practice: Reflections of an academic clinician. In E. F. Bourg, R. J. Bent, J. E. Callan, N. F. Jones, J. McHolland & G. Stricker (Eds.), *Standards and evaluation in the education and training of professional psychologists*. Norman, OK: Transcript Press. Pp. 7-23.

Meehl, P. E. (1989a) Psychological determinism or chance: configural cerebral autoselection as a tertium quid. In M. L. Maxwell & C. W. Savage (Eds.), *Science, mind, and psychology: Essays in honor of Grover Maxwell*. Lanham, MD: Univer. Press of America. Pp. 211–255. [Reprinted in Meehl, *Selected philosophical and methodological papers* (C. A. Anderson & K. Gunderson, Eds.). Minneapolis, MN: Univer. of Minnesota Press, 1991. Pp. 136-168.]

Meehl, P. E. (1989b) Schizotaxia revisited. *Archives of General Psychiatry*, 46, 935-944.
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. Reply to the commentators, pp. 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) Schizotaxia as an open concept. In A. I. Rabin, R. Zucker, R. Emmons, & S. Frank (Eds.), *Studying persons and lives*. New York: Springer. Pp. 248-303.
Meehl, P. E. (1990d) Toward an integrated theory of schizotaxia, schizotypy and schizophrenia. *Journal of Personality Disorders*, 4, 1-99.
Meehl, P. E. (1990e) Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, 66, 195-244. [Also see 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. (1991/1993) Four queries about factor reality. *History and Philosophy of Psychology Bulletin*, 5(No. 2), 4-5. [Initial 1991 publication contained erroneous title and other errors; this reference is to corrected publication]
Meehl, P. E. (1992a) Factors and taxa, traits and types, differences of degree and differences of kind. *Journal of Personality*, 60, 117-174.
Meehl, P. E. (1992b) The Miracle Argument for realism: an important lesson to be learned by generalizing from Carrier's counter-examples. *Studies in History and Philosophy of Science*, 23, 267-282.
Meehl, P. E. (1992c) Needs (Murray, 1938) and state-variables (Skinner, 1938). *Psychological Reports*, 70, 407-450.
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., Lykken, D. T., Schofield, W., & Tellegen, A. (1971) Recaptured-item technique (RIT): a method for reducing somewhat the subjective element in factor-naming. *Journal of Experimental Research in Personality*, 5, 171-190.
Merton, R. K. (1973) *The sociology of science*. Chicago: Univer. of Chicago Press.
Miller, D. (1972) The truth-likeness of truthlikeness. *Analysis*, 33, 50–55.
Millikan, R. A. (1917) *The electron*. Chicago, IL: Univer. of Chicago Press, 1963. (Reprint)
Mitroff, I. (1974) *The subjective side of science*. Amsterdam: Elsevier.
Murphy, N. (1989) Another look at novel facts. *Studies in History and Philosophy of Science*, 20, 385–388.
Murray, H. A. (1938) *Explorations in personality*. New York: Oxford Univer. Press.
Mynatt, C. R., Doherty, M. E., & Tweney, R. D. (1977) Confirmation bias in a simulated research environment: an experimental study of scientific inference. *Quarterly Journal of Experimental Psychology*, 29, 85-95.
Mynatt, C. R., Doherty, M. E., & Tweney, R. D. (1978) Consequences of confirmation and disconfirmation in a simulated research environment. *Quarterly Journal of Experimental Psychology*, 30, 395-406.
Nagel, E. (1939) Principles of the theory of probability. *International Encyclopedia of Unified Science*, 1(6).
Newton-Smith, W. H. (1981) *The rationality of science*. Boston, MA: Routledge & Kegan Paul.
Nickles, T. (1986) Remarks on the use of history as evidence. *Synthese*, 69, 253-266.
Niiniluoto, I. (1984) *Is science progressive?* Boston, MA: D. Reidel.
Niiniluoto, I. (1987) *Truthlikeness*. Boston: D. Reidel.
Niiniluoto, I. (1991) Discussion: Goldstick and O'Neill on "truer than." *Philosophy of Science*, 58, 491-495.
Nisbett, R. E., Fong, G. T., Lehman, D. R., & Cheng, P. W. (1987) Teaching reasoning. *Science*, 238, 625-631.
Nisbett, R. E., & Ross, L. (1980) *Human inference: strategies and shortcomings of human judgment*. Englewood Cliffs: Prentice-Hall.

Nozick, R. (1974) *Anarchy, state and utopia*. New York: Basic Books.
Oddie, G. (1986) *Likeness to truth*. Boston, MA: D. Reidel.
Oddie, G. (1990) Verisimilitude by power relations. *British Journal for the Philosophy of Science*, 41, 129-146.
O'Hear, A. (1980) *Karl Popper*. Boston: Routledge & Kegan Paul.
Oldroyd, D. (1986) *The arch of knowledge: an introductory study of the history of the philosophy and methodology of science*. New York: Methuen.
Overbye, D. (1991) *Lonely hearts of the cosmos*. New York: HarperCollins.
Peirce, C. S. (1878) How to make our ideas clear. In C. J. W. Kloesel (Ed.), *Writings of Charles S. Peirce*. Vol. 3. Bloomington, IN: Indiana Univer. Press, 1986. Pp. 257-276. [Originally published in *Popular Science Monthly*, 12, 286-302.]
Popper, K. R. (1959) *The logic of scientific discovery*. New York: Basic Books. (Original work published 1935)
Popper, K. R. (1962) *Conjectures and refutations*. New York: Basic Books.
Popper, K. R. (1972) *Objective knowledge*. Oxford, Eng.: Clarendon.
Popper, K. R. (1976) A note on verisimilitude. *British Journal for the Philosophy of Science*, 27, 147–195.
Popper, K. R. (1983) *Postscript* (Vol. I): *Realism and the aim of science*. Totowa, NJ: Rowman & Littlefield.
Pound, R. (1959) *Jurisprudence*. St. Paul, MN: West.
Price, D. J. de S. (1963) *Little science, big science*. New York: Columbia Univer. Press.
Rabb, T. K. (1983) The development of quantification in historical research. *Journal of Interdisciplinary History*, 13, 591-601.
Reichenbach, H. (1938) *Experience and prediction*. Chicago: Univer. of Chicago Press.
Richardson, M. W. (1941) The combination of measures. In P. Horst, *Prediction of personal adjustment*. New York: Social Sciences Research Council, Bulletin No. 48. Pp. 379-401.
Romanes, G. J. (1883) *Animal intelligence*. New York: Appleton.
Rousseau, D. L. (1992) Case studies in pathological science. *American Scientist*, 80, 54-63.
Rowney, D. K., & Graham, J. Q., Jr. (Eds.) (1969) *Quantitative history: selected readings in the quantitative analysis of historical data*. Homewood, IL: Dorsey.
Russell, B. (1948) *Human knowledge, its scope and limits*. New York: Simon & Schuster.
Salmon, W. C. (1984) *Scientific explanation and the causal structure of the world*. Princeton, NJ: Princeton Univer. Press.
Santayana, G. (1923) *Scepticism and animal faith*. New York: Scribner's.
Sawyer, J. (1966) Measurement and prediction, clinical and statistical. *Psychological Bulletin*, 66, 178-200.
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.
Schilpp, P. A. (Ed.) (1974) *The philosophy of Karl Popper*. LaSalle, IL: Open Court.
Schlesinger, A. M., Jr. (1962) The humanist looks at empirical social research. *American Sociological Review*, 27, 768-771.
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.
Shimony, A. (1955) Coherence and the axioms of confirmation. *Journal of Symbolic Logic*, 20, 1-28.
Shimony, A. (1967) Amplifying personal probability theory: comments on L. J. Savage's "Difficulties in the theory of personal probability." *Philosophy of Science*, 34, 326-332.
Siegel, H. (1980) Justification, discovery and the naturalizing of epistemology. *Philosophy of Science*, 47, 297-321.
Simonton, D. K. (1990) *Psychology, science, and history: an introduction to historiometry*. New Haven, CT: Yale Univer. Press.
Sines, J. O. (1970) Actuarial versus clinical prediction in psychopathology. *British Journal of Psychiatry*, 116, 129-144.

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

Skinner, B. F. (1945) The operational analysis of psychological terms. *Psychological Review*, 54, 270–277, 291–294. [Reprinted in B. F. Skinner, *Cumulative record*. New York: Appleton-Century-Crofts, 1959. Pp. 272–286. Reprinted (slightly revised) *Behavioral and Brain Sciences*, 1984, 7, 547-553, Author's response [to commentaries], pp. 572-581. Reprinted (with Postscript) in A. C. Catania & S. Harnad (Eds.), *The selection of behavior: the operant behaviorism of B. F. Skinner: comments and consequences*. New York: Cambridge Univer. Press, 1988. Pp. 150-164.]

Slovic, P., & Fischoff, B. (1977) On the psychology of experimental surprises. *Journal of Experimental Psychology, Human Perception and Performance*, 3, 544-551.

Smith, M. L., Glass, G. V., & Miller, T. I. (1980) *The benefits of psychotherapy*. Baltimore, MD: Johns Hopkins Univer. Press.

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

Sneed, J. D. (1979) *The logical structure of mathematical physics*. (2d ed.) Boston, MA: D. Reidel.

Sternbach, R. A. (1963) Congenital insensitivity to pain: a critique. *Psychological Bulletin*, 60, 252-264.

Stone, L. (1979) The revival of narrative: reflections on a new old history. Past and Present, 85, 3-24. [Reprinted in *The past and the present*. Boston, MA: Routledge & Kegan Paul, 1981. Pp. 74-96.]

Sulloway, F. J. (1990) Orthodoxy and innovation in science: the influence of birth order in a multivariate context. Paper presented at the meeting of the American Association for the Advancement of Science, New Orleans, LA, February 16, 1990.

Tarski, A. (1944) The semantic conception of truth and the foundations of semantics. *Philosophy and Phenomenological Research*, 4, 341-375. [Reprinted in H. Feigl & W. Sellars (Eds.), *Readings in philosophical analysis*. New York: Appleton-Century-Crofts, 1949. Pp. 52-84.]

Tarski, A. (1956) The concept of truth in formalized languages. In A. Tarski, *Logic, semantics, metamathematics*. Oxford: Clarendon. Pp. 152-278. [Originally published in German, 1935]

Tichý, P. (1978) Verisimilitude revisited. *Synthese*, 38, 175–196.

Tilly, C. (1984) The old new social history and the new old social history. *Review*, 7, 363-406.

Tolman, E. C. (1932) *Purposive behavior in animals and men*. New York: Century.

Tuomela, R. (Ed.) (1978) Verisimilitude and theory-distance. *Synthese*, 38, 213–246.

Tweney, R. D., Doherty, M. E., & Mynatt, C. R. (Eds.) (1981) *On scientific thinking*. New York: Columbia Univer. Press.

Venn, J. (1888) *The logic of chance*. New York: Macmillan.

Von Mises, R. (1939) *Probability, statistics and truth*. New York: Macmillan.

Wachter, K. W., Hammel, E. A., & Laslett, P. (1978) *Statistical studies of historical social structure*. New York: Academic Press.

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

Wiggins, J. S. (1981) Clinical and statistical prediction: where are we and where do we go from here? *Clinical Psychology Review*, 1, 3-18.

Wilks, S. S. (1938) Weighting systems for linear functions of correlated variables when there is no dependent variable. *Psychometrika*, 3, 23-40.

Wittgenstein, L. (1922) *Tractatus logico-philosophicus*. New York: Harcourt, Brace.

Worrall J. (1985) Scientific discovery and theory-confirmation. In J. C. Pitt (Ed.), *Change and progress in modern science: papers related to and arising from the Fourth International Conference on History and Philosophy of Science*. Dordrecht: Reidel. Pp. 301-331.

Worrall J. (1989) Fresnel, Poisson and the white spot: the role of successful predictions in the acceptance of scientific theories. In D. Gooding, T. Pinch, & S. Schaffer (Eds.), *The use of experiment: studies in the natural sciences*. Cambridge, Eng.: Cambridge Univer. Press. Pp. 135-157.

Accepted July 8, 1992.

APPENDIX 1

INSENSITIVITY OF A LINEAR VERISIMILITUDE COMPOSITE TO VARIATION IN WEIGHTING 10 *V*-LEVELS

Consider a set of *m*-Postulate theories each appraised as to its agreement with the textbook-accepted criterion theory, here treated as quasi-gold standard. Each of the *m* postulates of a theory can "pass" or "fail" a level of 10 specification levels I–X. The number of its postulates passing a level is the theory's *level score*. Thus each theory has 10 level scores, x_1 = number of postulates passing level I, x_2 = number of postulates passing level II, …, x_{10} = number of postulates passing level X. The levels being defined as Guttman scalable, if a postulate "fails" at level k, it also fails at level $(k + 1)$ and at all higher levels. Hence the theory's 10 scores are everywhere non-increasing, $x_1 \geq x_2 \geq x_3 \geq \ldots\ x_9 \geq x_{10}$. We construct a verisimilitude index as a linear composite of the 10 level scores. I denote any composite verisimilitude index, however computed from level scores (e.g., the optimal function, in Utopian metatheory, may not be linear), by the handy term ⌜Vindex⌝, despite that being a Roman general's name. How shall the 10 scores be weighted? This Appendix shows that, for any but the most extreme, unplausible, and pessimistic conditions, the weighting choice will make negligible difference even in the high correlation region $r_{VT} > .95$ (between verisimilitude and track record).

We use Burt's formula (Burt, 1950)

$$r = 1 - \frac{\left(1 - \overline{r}\right)}{2n\overline{r}}\left(\frac{\sigma_{w_1}^2}{\overline{W}_1^{\,2}} + \frac{\sigma_{w_2}^2}{\overline{W}_2^{\,2}}\right) \qquad [1]$$

for the expected correlation between two randomly weighted composites, where $\overline{r}$ = mean pairwise correlation of the variables, n = number of variables in the composite, σ_{W1} and σ_{W2} the standard deviations of the weights in the 2 weighting systems, $\overline{W}_1$ and $\overline{W}_2$ the mean weights in the two systems. Cronbach (1951) considers Burt (1950) and Richardson (1941) to have given the clearest mathematical analyses of the differential weighting problem.

The probability p_k of a postulate passing a level k is the proportion [= empirical relative frequency] that pass it over all postulates and all theories. To simplify at first, I shall assume that p_k is constant and equals p over levels and that whether a given postulate in a theory passes a level is independent of whether another postulate passes. (Obviously there is no rational basis to identify postulates over different theories. The *m* postulates are not "numbered" except in the nominal use of numerals.) To apply Burt's formula we must estimate the average interlevel correlation $\overline{r}$ among all level pairs $x_1\ x_2$, $x_1\ x_3$, $x_1\ x_4$, … $x_1\ x_{10}$, $x_2\ x_3$, … $x_i\ x_j$ … $x_9\ x_{10}$. There are $\binom{10}{2} = 45$ interlevel score

correlations. (NB: The *source of individual differences* giving rise to these correlations is the *variation over theories* in each of their level scores.)

Our first task is to derive a formula for the correlation between adjacent level scores as a function of the interlevel transitional probability p, which I shall assume is also the individual probability of a postulate "passing" level I. (These strong, improbable simplifying assumptions will be relaxed later.) Then the probability of a postulate passing level $\mathrm{I} = p$, of passing levels $\mathrm{I} + \mathrm{II} = p^2$, … of passing all 10 levels $= p^{10}$. Assuming independence, the expected value of a theory's score at level k is $m\,p^k$, and the variance of theory scores at that level is $m\,p^k\,(1 - p^k)$. For theories having score x_k at level k, the expected value $\hat{x}_{k+1}$ of their level $(k + 1)$ score is $p\,x_k$. The mean level scores thus being proportional to the preceding level scores, the stochastic dependence of x_{k+1} on x_k is linear. Hence we may write the correlation between level scores using the usual slope formula (r being the slope of the best fitting straight line when both variables are standardized). Call $\sigma_{x_k} = \sigma_k$ for short, similarly $\sigma_{x_{k+1}} = \sigma_{k+1}$. Then the slope formula predicting x_{k+1} from x_k is

$$\frac{\hat{x}_{(k+1)\cdot k} - \overline{x}_{k+1}}{\sigma_{k+1}} = r_{k(k+1)}\left(\frac{x_k - \overline{x}_k}{\sigma_k}\right) \qquad [2]$$

$$\sigma_k = \left[mp^k\left(1 - p^k\right)\right]^{1/2} \qquad [3]$$

$$\sigma_{k+1} = \left[mp^{k+1}\left(1 - p^{k+1}\right)\right]^{1/2} \qquad [4]$$

$$\overline{x}_k = mp_k = mp^k \qquad [5]$$

$$\overline{x}_{k+1} = mp_{k+1} = mp^{k+1} \quad . \qquad [6]$$

Expressing r in terms of these variables

$$r_{k(k+1)} = \frac{\hat{x}_{k(k+1)\cdot k} - \overline{x}_{k+1}}{\left[mp^{k+1}\left(1 - p^{k+1}\right)\right]^{1/2}} \Big/ \frac{x_k - \overline{x}_k}{\left[mp^k\left(1 - p^k\right)\right]^{1/2}}$$

$$= \left(\frac{x_k p - mp^{k+1}}{x_k - mp^k}\right)\left(\frac{mp^k\left(1 - p^k\right)}{mp^{k+1}\left(1 - p^{k+1}\right)}\right)^{1/2} \qquad [7]$$

$$= p^{1/2}\left(\frac{1 - p^k}{1 - p^{k+1}}\right)^{1/2} \qquad [8]$$

which fits the intuition that the correlation between level scores is an increasing function of the level-to-level transitional probability p.

Analogous derivations (here omitted) for the correlation between non-adjacent levels k and $k + 2$ separated by two steps yields

$$r_{k(k+2)} = p\left(\frac{1-p^{k}}{1-p^{k+2}}\right)^{1/2} \qquad [9]$$

and for levels separated by three steps

$$r_{k(k+3)} = p^{3/2}\left(\frac{1-p^{k}}{1-p^{k+3}}\right)^{1/2} \qquad [10]$$

and so on for any number of steps, the extra-bracket p having an exponent that increases by 1/2 power with each intervening step. Since $p < 1$, these increasing powers work to reduce inter-level correlations as more levels lie between the two being correlated. Each non-adjacent formula is the product of the intervening adjacent formulas (proof omitted), so we have the relation (where s = number of steps)

$$r_{k(k+\mathrm{s})} = r_{k(k+1)} \cdot r_{(k+1)(k+2)} \cdot \ldots \cdot r_{(k+\mathrm{s}-1)(k+\mathrm{s})} \quad . \qquad [11]$$

This fits the intuition that when correlations are generated by a sequence of linear operations on scores $\bar{x}_{(k+1)\cdot x_k} = px_k$ for adjacent levels, the successive linear dependencies (slopes) can be "piggy-backed," as in path analysis.

It is easily shown (proof omitted) that the adjacent level rs increase as we move through the levels from I to X. Therefore, assuming them all to be equal to the first one would underestimate their sum. Making that counterfactual assumption,

$$r_{13} = r_{12}r_{23} = r_{12}^{2}$$

$$r_{14} = r_{12}r_{23}r_{34} = r_{12}^{3}$$

$$\vdots$$

$$r_{ij} = r_{i(i+1)}r_{(i+1)(i+2)}\cdots r_{i(i+\mathrm{s})} = r_{i+1}^{\mathrm{s}}$$

the sum of interlevel correlations for a given level with all higher levels is a geometric series in r,

$$\mathrm{S}_{n} = \frac{a\left(1-r^{n}\right)}{1-r}$$

$$= \frac{r\left(1-r^{n}\right)}{1-r} \qquad \text{since initial term } a = r. \quad [12]$$

Summarizing this over all 45 pairs of levels

$$\sum \mathrm{S}_{n} = \frac{r\left(1-r^{9}\right)}{1-r} + \frac{r\left(1-r^{8}\right)}{1-r} + \ldots + \frac{r\left(1-r\right)}{r}$$

$$= \frac{r}{1-r}\left(9 - \sum_{i=1}^{k} r^{k}\right) \quad . \qquad [13]$$

But the inner sum is another geometric series identical with the first one, so we obtain

$$\sum \mathrm{S}_{n} = \frac{r}{1-r}\left(9 - \frac{r\left(1-r^{9}\right)}{1-r}\right) \qquad [14]$$

and dividing by $\binom{10}{2} = 45$ to get the mean of these 45 correlations

$$\overline{r} = \frac{1}{45}\left(\frac{r}{1-r}\right)\left(9 - \frac{r\left(1-r^{9}\right)}{1-r}\right) \qquad [15]$$

which underestimates $\overline{r}$ by some unknown amount.

We need a "safe" (too low) value for p to plug into this formula. I conjecture that the probability of a single postulate in a scientific theory being correct is not smaller than .10, allowing a reasonable numerical tolerance for the parameters in level X.[45] Given a single postulate's correctness probability (passing levels I–X) as $p_{\mathrm{X}} = .10$, then the initial and transitional probability $p = (.10)^{1/10} = .79$, which put into Equation 15 yields $\overline{r} = .48$ as a safe lower bound for use in Burt's formula.

Suppose the 10 weights in a weighting system were distributed as in the frequency polygon of $(p+q)^{10}$ for $p = q = ½$, rounding to integer counts. This gives weights ranging from 3 to 8 with a mean $\overline{\mathrm{W}} = 5.50$ and $SD = 1.5$, so $\sigma_{\mathrm{W}}^{2} / \overline{\mathrm{W}}^{2} = .0744$. Let both of the systems have this weight distribution (the weights being

[45]Plausibility argument: Assuming that value, if p-independence across postulates obtained, a 10-postulate theory would have only 1 chance in a septillion of being correct, which seems rather pessimistic. Or, taking account of the patterned purposive character of theorizing, assume that 10-postulate theories pass or fail verisimilitude levels in organized blocks, say, the 3, 3, and 4 postulates in each block being *completely linked* as to verisimilitude—an absurdly strong assumption—and the 3 block probabilities being $p = .10$. Even then, a theory has only 1 chance in 1000 of being correct, still pretty pessimistic.

uncorrelated as Burt assumes). Then the correlation between Vindex_1 and Vindex_2 is

$$r_{V_1V_2} = 1 - \frac{(1-.48)}{10(.48)}(.0744) = .992 \quad .$$

Thus two scientific realists could assign weights to the I–X *V*-levels thus distributed *but uncorrelated*, and their verisimilitude appraisals would agree almost perfectly.

A more pessimistic scenario would be a rectangular distribution of weights one integer apart (1, 2, 3, … 10) by one evaluator, the other's being similar but with the assignments scrambled. For example, in one system the "most important" level would get a weight 10 times as heavy as the least important, twice heavier than the "middling important," and the middling weighted 5 times the least—a rather extreme dispersion of perceived "importance" over levels. The other evaluator assigns a similarly distributed set, but uncorrelated with the first one. Thus, one evaluator may weight level I, the correct kind of theoretical entity, as 10, but the other one gives it a low or middling weight. Using the formulas for the sum of integers from 1 to 10, and sum of their squares, we obtain $\sigma_W^2 / \overline{W}^2 = .27$, which put into Burt's formula gives $r = .97$.

A plausible conjecture is that the initial and transitional *p*-values for postulates "passing" *V*-levels will vary over levels. The formula (proof omitted, proceeds analogously to the fixed *p* case) for the Pearson *r* between scores on adjacent levels *k*, $k + 1$ is

$$r_{k(k+1)} = p_{(k+1)/k}^{1/2} \left(\frac{1-p_k}{1-p_{(k+1)}} \right)^{1/2}$$

that is, the specific inter-level transitional probability replaces the former generic *p*. The formula for a nonadjacent level correlation between levels *k*, $k + 2$ (proof omitted, analogous) is

$$r_{(k+2)k} = \left(p_{(k+2)/(k+1)} \cdot p_{(k+1)/k} \right)^{1/2} \left(\frac{1-p_k}{1-p_{(k+2)/(k+1)} \cdot p_{(k+1)/k}} \right)^{1/2} \quad .$$

In general, for nonadjacent levels,

$$r_{k(k-s)} = \left(p_{k/(k+1)} \cdot p_{(k+1)/(k+2)} \cdots p_{(k+s-1)/(k+s)} \right)^{\frac{s-1}{2}} \times \left(\frac{1-p_k}{1-p_k p_{(k+1)/k} p_{(k+2)/(k+1)} \cdots p_{(k+s)/(k+s-1)}} \right)^{1/2} \quad .$$

Again, as in the fixed p case, nonadjacent level correlations can also be computed directly from the adjacent by piggy-backing.

All of these formulas have been "checked empirically" by Monte Carlo runs, where the table of 45 pairwise levels correlations is produced three ways: (a) By actually generating postulate scores atomistically from values of p, "from the ground up"; (b) By the formulas for r, adjacent and nonadjacent; (c) By multiplying adjacent rs to get the nonadjacent ("piggy-back" method). The tables agree within sampling and rounding error, the second-order correlations between tabled rs [likewise z_rs] computed three ways being uniformly > .99.

I have conducted a variety of Monte Carlo runs assigning different parameter configurations (transitional probabilities, range, and correlation between weighting systems) which will not be reported in detail here. The clear finding is, as anticipated from the research on weighting composites cited in Meehl (1990b, p.19), that in our 10-variable system with positive manifold, the linear Vindex composite is highly insensitive to changes in the weights.

Example: Assume a pass-probability $p = .63$ to generate level scores. Two metatheorists assign subjective, intuitive "importance" similarly on the top five levels and bottom level, but completely reverse their weights on the other four thus:

	Level									
	I	II	III	IV	V	VI	VII	VIII	IX	X
w_1:	10	9	8	7	6	5	4	3	2	1
w_2:	10	9	8	7	6	2	3	4	5	1

The two Vindex composites correlate $r = .999$.

Example: Given a pass-probability of $p = .50$ and two rectangular systems of weights correlating .03 with each other, the resulting Vindex composites correlate $r = .99$.

Example: If $p = .50$ but the two rectangular weighting systems are completely reversed, one still obtains $r_{W_1W_2} = .86$ between the Vindex composites.

Example: Let the initial and transitional probabilities range from level I $p = .95$ to last transitional $p_{IX-X} = .35$, and assign rectangular weighting systems correlating .03 with one another. Then the Vindex composites correlate .93.

Of course, we have no basis for thinking that any two "rational" Vindex weight sets would correlate zero, let alone negatively. An empirically selected weighting system, whether based on internal relations (factor analysis, taxometrics, multidimensional scaling) or some reasonable external "criterion" (experimental track record) would presumably tend to agree fairly well with another so based. As suggested in the text,

absent strong intuitive value preferences as to the 10 levels' "importance" (to a scientific realist), canonical correlation between *V*-level scores and performance measures seems a reasonable way to proceed. If a scientific realist complains that the Vindex obtained thus, or by assigning equal weights, is strongly counterintuitive, the best (logical and pedagogical) reply would be to exhibit the sort of findings on scientist intuition reported in Appendix 2 (pp. 463-467).

APPENDIX 2

INTUITIVE WEIGHTING OF VERISIMILITUDE LEVELS' IMPORTANCE: LACK OF CONSENSUS IN A SAMPLE OF PSYCHOLOGISTS

Any doubts I had as to the importance of solving (or "dissolving") the Vindex weighting problem were dispelled by a pilot study of colleagues' opinions, which yielded a surprising lack of agreement among a set of highly competent scientists. I had anticipated sizable but imperfect pairwise correlations ($.50 < r < .70$), there being no "objective" criterion of importance and scholars differing as to how realist (versus fictionist, pragmatist, positivist, or operationist) they are in thinking about science; and yet (I thought) surely the first levels [getting the right *kind* of entities at least (level I) and their causal or compositional relations (level II)] would receive more emphasis than, say, the interaction effects (level VI) or the specific numerical values of function parameters (level X). I was in for a big surprise.

I selected a small group of colleagues ($N = 14$) of *excellent*—not just good—scientific competence, researchers whose publications are often cited, recipients of prizes, awards, offices in professional societies, and the like. Several of them could be properly called "world renowned," "eminent." But since these "social impact" criteria can sometimes mislead, I added to the community of scholars' consensus my own assessment, based on intellectual exchanges with these people over several years. They are all, in my judgment, not merely highly visible paper producers, but clearly first-class intellects. While they vary in technical knowledge of (and interest in) philosophy of science, I can attest that they are all methodologically sophisticated and can easily be drawn into discussion of meta-issues if one starts with a substantive problem. I added myself as a rater to these 14 to get $N = 15$ judges. Counting myself, four have participated as lecturers in my philosophical psychology seminar, and three have sat through it. Eight have published papers mainly methodological in content (two books). Fields of psychology represented (several multiples) were behavior genetics, clinical, differential, experimental (audition, comparative, emotional conditioning, learning, memory, psychophysiology, vision), neuroscience, and personality theory. All this by way of assuming that a more "qualified" sample of psychologists for the present purpose would be difficult to find.

Each was sent a memorandum and rating form two weeks prior to going on the payroll fall quarter 1991. After a lapse of four weeks, only six forms (including my own) having been returned, I followed up with a second memorandum.

This elicited no additional responses to the initial form. Of the nine who declined to make ratings, two said they did not understand verisimilitude as I explained it, one saw no clear basis for weighting so suggested equal weights, two checked external criterion

FIRST MEMORANDUM

TO: Selected Colleagues

FROM: P. E. Meehl

DATE: September 3, 1991

RE: Relative importance of aspects of theory verisimilitude

[In doing this task, you will help a lot with a paper I'm writing, "Cliometric metatheory." Spend only about 5 minutes on it, no more—if more, you're thinking too hard about it. Don't do it unless you know calculus.]

Note the attached "specifications" are increasingly detailed, and almost Guttman scalable. For example, a theory can't have the correct signs of derivatives of functions relating two theoretical entities if it doesn't relate them (III depends on II); nor the right function forms (VII) if it doesn't even get the derivative signs correct (III, IV).

If one theory is that held by Omniscient Jones [= the literally true theory = T_{OJ}], then the verisimilitude of a second theory T_i is its similitude to T_{OJ} in these 10 respects (attached list). We are considering the objective truth of the theory (ontology), not the evidence for it (epistemology).

Despite the quasi-scalability (don't fret that, not relevant to task) of the levels, scientists may differ in how many brownie points they give a theory for "passing" a given level. So although Level IV presupposes Level III, one may still ask how much more "important" one level is than the other. Someone might assign 2 points for passing Level IV and 3 points for passing Level V, despite the latter presupposing the former. (If that seems absurd to you, ok—then you won't do that.) Use your own intuition (or reason?) as to importance. I think of importance as how much I *care* about knowing such-and-such, or how thrilled I would be if *I* discovered the answer.

TASK: Assign weights in the range 1–10 (larger numbers = heavier weight = more "important"). You may distribute them as you please. For example, if you perceive only 2 degrees of value, your weights might read 2, 2, 1, 2, 1, …. These are *weights, not ranks*. I would prefer more dispersion, if possible.

If it seems a dumb thing to do, wait for the article, in which (as we say in paranoia) "Everything becomes clear."

RATING FORM

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

WEIGHTS: Use integers chosen from the interval [1, 10].
A higher number = more weight = more important.
Equal weights for one or more levels are possible.

YOUR FIELD(S) _____

_____

(Thank you. If you wish to know your correlation with others, sign your name or concoct a code number you will recognize.)

FOLLOW-UP MEMORANDUM

TO: Selected Colleagues

FROM: P.E.Meehl

DATE: September 30, 1991

RE: Verisimilitude ratings (PEM memo 9/3/91)

Since only 7 (of 15) have responded, which they did quickly, I suspect some of you find the task un-do-able or excessively onerous (despite my "five minutes" adjuration). Among the responders to date, what's striking is their poor agreement, contrary to my expectations. This has been illuminating and helpful to the project I'm engaged in.

If you find that you cannot do it without undue time or anxiety, please do me the favor of checking one or more of the following, which should be quick and easy:

_____ I don't understand verisimilitude as Meehl explains it.

_____ I think I understand verisimilitude but I reject the concept.

_____ I see no clear basis for weighting, so as far as I am concerned you could just as well weight them equally.

_____ They should be weighted on the basis of correlation with an external criterion, e.g., theory's track record of prediction and control (correlation between the 10 scores over a big batch of actual theories from history of science).

_____ Since Meehl claims they are Guttman scaleable, score a large batch of theories (where the answers are considered as long "settled" in standard textbooks) on the 10 aspects, factor analyze them, and weight by their correlation with the first big factor.

_____ A composite of such heterogeneous "apples and oranges" makes little or no sense to me, so why assign weights?

_____ Other reason: _____

correlation as an appropriate basis for weightings, and one also checked factor analysis. Two claimed their knowledge of calculus was inadequate or too rusty. One said he "didn't understand the task well enough to be able to do it." Two never replied, orally or in writing, to either note. Since the last three are bright, conscientious, cooperative high achievers as well as personal friends, my inference is that they found the task undoable (with any confidence) but did not want to say so.

Among the six of us who did rate (four with special interest and some formal education in philosophy of science), the 15 pairwise correlations range from –.92 to +.83 with a mean *r* precisely zero. Whatever one can say about Meehl's Verisimilitude Levels, it seems pretty clear that scientists' intuitions as to their relative importance are an unsafe guide! Suppose one had no external correlates (e.g., factual track record, "instrumental" composite employed in canonical correlation), and a factor analysis of interlevel correlations revealed no large, interpretable first factor; then the rational procedure would be to weight equally, since the expected value of the correlation between any pair of differentially weighted composites is less than the expected value of either with an equal weight composite (Dawes, 1970).

Although astonished (and, on reflection, still puzzled) by these findings, I do not think they speak against the Strong Actuarial Thesis. On the contrary, if anything they provide further ammunition for the Faust-Meehl doctrine. Except for level I (which requires some spelling out) and levels VIII and X (where conventional numerical tolerances to label parameters "correct" must be set), the verisimilitude components all have a precise meaning. Given a sample of actual theories from history of science, "scores" on the levels will, as a matter of *ascertainable fact*, exhibit such-and-such pairwise correlations. Multiple indexes of theory "performance" will be empirically related to each of the 10 *V*-scores and to a Vindex composite (however chosen) in certain ways, linear or otherwise. The only way to find out about these various correlational patterns is to sample episodes from the history of science, compute the actuarial values, and subject these to appropriate mathematical treatment. Should a scientist or metatheorist press for his favorite set of intuitive *V*-weights, we reply simply: (a) Intuition here is highly subjective, so why should we trust *his* in particular? (b) His proposed composite does a poorer job than one based on canonical correlation (as it certainly will) and no better than equal weights (as it probably will, unless he is a lucky genius); and (c) the weights do not matter much anyway, except in the *very* high correlation region, which his composite will not be in. Let me suggest "therapeutically," empirical research on both intra-verisimilitude patterns and performance correlates will undoubtedly provide insight into *why* some Vindex composites excel others, leading in turn to alteration in our metatheoretical intuitions—the usual path of scientific progress in reforming our common sense and pre-analytic conceptions.