Psychological Inquiry
1990, Vol. 1, No. 2, 108-141

#147

# Appraising and Amending Theories: The Strategy of Lakatosian Defense and Two Principles That Warrant It

**Paul E. Meehl**
*University of Minnesota*

*In social science, everything is somewhat correlated with everything ("crud factor"), so whether $H_0$ is refuted depends solely on statistical power. In psychology, the directional counternull of interest, H*, is not equivalent to the substantive theory T, there being many plausible alternative explanations of a mere directional trend (weak use of significance tests). Testing against a predicted point value (the strong use of significant tests) can discorroborate T by refuting H*. If used thus to abandon T forthwith, it is too strong, not allowing for theoretical verisimilitude as distinguished from truth. Defense and amendment of an apparently falsified T are appropriate strategies only when T has accumulated a good track record ("money in the bank") by making successful or near-miss predictions of low prior probability (Salmon's "damn strange coincidences"). Two rough indexes are proposed for numerifying the track record, by considering jointly how intolerant (risky) and how close (accurate) are its predictions.*

For almost three quarters of a century, the received doctrine about appraising psychological theories has been to perform a statistical significance test. In the "soft" areas (clinical, counseling, developmental, personality, and social psychology), where the primitive state of theory only rarely permits strong conjectures as to the mathematical functions (let alone their parameters!), refutation of the null hypothesis has usually been the sole theory-testing procedure employed. In the 1960s, several psychologists (Bakan, 1966; Lykken, 1968; Meehl, 1967; Rozeboom, 1960) came independently, for somewhat different reasons and hence with varied emphases, to entertain doubts as to the merits of null-hypothesis testing as a *theoretical* tool. (I set aside in this article the reliance on statistical significance in technology—e.g., benefit of a psychotropic drug, efficacy of an instructional method.) At the close of that decade, sociologists Morrison and Henkel (1970) edited a volume reprinting critical articles, and replies to them, by biologists, sociologists, psychologists, statisticians, and an economist. This excellent book should by rights be called "epoch-making" or "path-breaking," but, regrettably, it was not. I do not know how well it sold, but it is rarely cited; and I find that the majority of psychology students in my department have never heard of it, let alone been urged by their professors to read it. Judging from published research in current soft psychology, the PhD orals I serve on, and colloquium lectures by job candidates, the book has had negligible influence.

My first article on this topic (Meehl, 1967) focused on the paradox that improved instrumentation and sample size results in a stiffer test—greater danger of theory refutation—in physics, whereas the reverse is true in psychology. The reason for that lies in the way significance tests are normally used in the two disciplines. In physics, one typically compares the observed numerical value with the theoretically predicted one, so a significant difference refutes the theory. In social science, the theory being too weak to predict a numerical value, the difference examined is that between the observed value and a null ("chance") value, so statistical significance speaks *for* the theory. Holding the meta-theoretical views of Sir Karl Popper, I argued that this was an unhealthy state of affairs in that it did not provide the psychological researcher with strong ("risky," "dangerous," and hence highly corroborative) tests.

Ten years later, I wrote at greater length along similar lines (Meehl, 1978); but, despite my having received more than 1,000 reprint requests for that article in the first year after its appearance, I cannot discern that it had more impact on research habits in soft psychology than did Morrison and Henkel. Our graduate students typically plan and write their doctoral dissertations in blissful ignorance that "the significance test controversy" even exists, or could have a bearing on their research problems. This article (see also Meehl, 1990c) is my final attempt to call attention to a methodological problem of our field that I insist is not minor but of grave import.

I am incidentally replying to Serlin and Lapsley (1985), who advanced discussion of the issue by correcting my overly Popperian stance ("strict falsificationism") and pointing out that it is more realistic to think of theories as being good enough [even if, literally, false]" than to set up a rigid true/false dichotomy in the way I did in 1967 and 1978. I cheerfully accept their criticism, as well as their "good enough" principle, although I am not convinced that their specific *statistical* implementation of the principle is as helpful as they think. (This is not primarily a statistical disagreement, but one of methodological focus, as I shall argue at length.) A strong contribution by Dar (1987) advanced the discussion, but, because I agree with practically every sentence he wrote, I shall not consider him further. That Imre Lakatos (1970; Worrall & Currie, 1978a, 1978b) would disagree with Serlin and Lapsley's application of their "good enough" principle to most of social science theories (*and* experiments), I can attest from many hours of conversation with him. He viewed most social science pretty much as does Andreski (1972) and in conversation was even more contemptuous than in print, usually characterizing the books and articles as being harmful to our environment, a form of "intellectual pollution." In 1967 I had never heard of Lakatos, and I met him for the first time when he visited the Minnesota Center for Philosophy of Science some time in 1969 (Lakatos, in Worrall & Currie, 1978a, p. 87 fn. 3). As to Serlin and Lapsley's complaint that, although I cited him in my 1978 article, I did not integrate his views with my neo-Popperian critique of significance testing, the reasons for that were (a) space and (b) doubts as to whether I could do it. I think now that I can, but I'm not sure. Moving from Popper to Lakatos does *not* appreciably soften the blow of my 1967 attack, and here I shall try to show that a proper interpretation of Serlin and Lapsley's "good enough" principle must rely on two other principles, both Popperian in spirit although not "orthodox Popper."

### Theory Appraisal in Current Metatheory

To further discussion of the role of significance testing it is necessary to set out a general conception of theory appraisal in current metatheory, which I must do briefly and hence with an unavoidable flavor of dogmatism. Most of what I shall say is, I believe, fairly generally agreed on among philosophers of science. I prefer the term 'metatheory' for philosophy of science as currently understood—that is, theory of theories, the rational reconstruction of empirical history of science, eventuating in a mixed normative and descriptive content. More generally, scientific metatheory is a subdivision of what has come to be called "naturalized epistemology." The *prescriptive* component attempts to "advise" the scientist with guidelines or principles—not strict rules—derived from the *descriptive* findings of historians of science as to what has succeeded and what has failed, to the extent that success or failure reveals methodological trends. I could call the position '*neo*-Lakatosian', as the late Imre Lakatos might not agree with all of it. For ease of reference, I set out the general position with brief numbered paragraphs and minimum elaboration or defense.

1. A scientific theory is a set of statements in general form which are interconnected in the sense that they contain overlapping terms that designate the constructs of the theory. In the nomological network metaphor (Cronbach & Meehl, 1955), the nodes of the net are the theoretical constructs (entities) and the strands of the net are the functional or compositional laws relating them to one another. Contrary to simplistic operationism, it is not required that all the theoretical constructs be operationally defined. Only a proper subset are linked in a direct way to observational predicates or statements. In idealization, the theory consists of a formal calculus and an embedding text that provides the interpretation of expressions in the formalism (cf, Suppe, 1977). The empirical meaning of the theoretical terms is given partly by "upward seepage" from the subset that are operationally tied to the data base. Logicians explicate this upward seepage by means of a technical device called the Ramsey sentence, which eliminates the theoretical terms without "eliminating the theory" or repudiating its existence claims. For psychologists its importance lies more in showing (contrary to simplistic positivism and freshman rhetoric class) how a system of expressions can both *define* and *assert* concurrently. A clear and succinct exposition of the Rarnsey sentence can be found in Carnap (1966, chap. 26 and pp. 269-272). For additional discussion, see, in order Maxwell (1962, pp. 15ff; 1970, pp. 187-192), Glymour (1980, pp. 20-29), and Lewis (1970).

In addition to this "implicit definition by Ramsified upward seepage," empirical meaning of theoretical terms is contributed partly by an interpretive text that *characterizes* the theoretical entities and their relations in various ways. Sometimes this interpretive text does its job by reducing concepts to concepts lower in the pyramid of the sciences, but not always. There are some interesting generic terms that cut across disciplines, so that the appearance of these terms in the embedding text does not tell us what science we are pursuing. Examples are 'cause,' 'influence,' 'inhibit,' 'retard,' 'potentiate,' 'counteract,' 'form,' 'be composed of,' 'turn into,' 'interact with,' 'vanish,' 'link,' 'accelerate,' 'modify,' 'facilitate,' 'prevent,' 'change,' 'merge with,' 'produce,' 'adjoin,' 'converge upon,' and the like. I have doubts as to whether these interesting words, which perhaps an old-fashioned philosopher of science would have called metaphysical, and which occur in the interpretive text of such diverse sciences as economics, chemistry, behavior genetics, and paleontology with similar (sometimes identical) meaning, can be Ramsified out. But I have not seen any discussion of this in the metatheoretical literature. They are *not* metalinguistic terms, but are object language terms of a highly general nature.

2. In conducting an empirical test of a substantive theory $T$ (which it is imperative to distinguish from a test of the statistical hypothesis $H$) the logical form is the following:

$$(T \,.\, A_t \,.\, C_p \,.\, A_i \,.\, C_n) \rightarrow (O_1 \supset O_2)$$

where $T$ is the theory of interest, $A_t$ the conjunction of auxiliary theories needed to make the derivation to observations go through, $C_p$ is a *ceteris paribus* clause ("all other things being equal"), $A_i$ is an auxiliary theory regarding instruments, and $C_n$ is a statement about experimentally realized conditions (particulars). The arrow denotes deduction (entailment), and on the right is a material conditional (horseshoe) which says that if you observe $O_1$ you will observe $O_2$. ($O_1$ and $O_2$ are not, of course, related by strict entailment.) On careful examination one always finds in fields like psychology that the auxiliary $A_t$ is itself a conjunction of several auxiliary theories $A_1, A_2, \ldots, A_m$. If in the laboratory, or in our clinic files, or in our field study, we observe the conjunction ($O_1 \,.\, {\sim}O_2$) which falsifies the right-hand conditional, the left-hand conjunction is falsified *modus tollens* (Popper, 1935/1959, 1962; Schilpp, 1974; cf. O'Hear, 1980).

3. Although *modus tollens* is a valid figure of the implicative syllogism, the neatness of Popper's classic falsifiability concept is fuzzed up by the fact that negating the left-hand conjunction is logically equivalent to stating a disjunction of the negations, so that what we have achieved by our laboratory or correlational "falsification" is a falsification of the *combined* claims $T \,.\, A_t \,.\, C_p \,.\, A_i \,.\, C_p$, which is not what we had in mind when we did the experiment. What

happens next is therefore not a matter of formal logic, but of scientific strategy. All the logician can tell us here is that *if* we accept the observational conjunction ($O_1 \,.\, \sim O_2$), *then* we will necessarily deny the fivefold conjunction on the left (Meehl, 1978, 1990c).

4. If this falsification does not occur, we say that the theory has been *corroborated,* which for Popper means that it has been subjected to a test and has not failed it. Whatever affirmative meaning (reliance? "animal faith"? rational scientific belief?) we give to corroboration derives from a further statement, namely, that absent the theory *T,* the antecedent probability of $O_2$ conditional upon $O_1$ is "small." If that is not so, our corroboration (pre-Popperians called it *confirmation,* a term that Popper avoids as being justificationist) is weak, some say negligible. Because if we say that the left is proved because the right-hand side is empirically correct, this inference is formally invalid, being the fallacy of "affirming the consequent." The logicians' old joke here, attributed to Morris Raphael Cohen, makes the point: "All logic texts are divided into two parts. In the first part, on deductive logic, the fallacies are explained; in the second part, on inductive logic, they are committed." When we speak of the theory as "taking a risk," as "surmounting a high hurdle," as not being flunked "despite a dangerous test," these locutions refer to the notion that on some basis (prior experience, other theory, or common knowledge and intuition), *absent the theory T we have our eye on*, we see no reason for thinking that $O_2$ has a high probability conditional upon $O_1$.

5. The obvious way in which we warrant a belief that $O_2$ has a low prior probability conditional upon $O_1$ absent the theory is when $O_2$ refers to a point value, or narrowly defined numerical interval, selected from a wide range of otherwise conceivable values. The precise explication of this risky-test notion is still a matter of discussion among logicians and philosophers of science (cf. Giere, 1984, 1988) but I presuppose the basic idea in what follows. Because not all psychologists subscribe to a Popperian or Lakatosian metatheory, I must emphasize that one need not subscribe to Popper's anti-inductivism, nor to his emphasis on falsification, to accept the notion of risky test, perhaps expressed in other, less committed language. Working scientists who never heard of Popper, and who have no interest in philosophy of science, have for at least three centuries adopted the position that a theory predicting observations "in detail," "very specifically," or "very precisely" gains plausibility from its ability to do this. I have not met any scientist, in any field, who didn't think this way, whether or not he had ever heard of Karl Popper. If my meteorological theory successfully predicts that it will rain sometime next April, and that prediction pans out, the scientific community will not be much impressed. If my theory enables me to correctly predict which of 5 days in April it rains, they will be more impressed. And if I predict how many millimeters of rainfall there will be on each of these 5 days, they will begin to take my theory very seriously indeed. That is just scientific common sense, part of the post-Galilean empirical tradition that does not hinge on being a disciple of Popper or Lakatos.

6. By the instrumental auxiliaries $A_i$ I mean the accepted theory of devices of *control* (such as holding a stimulus variable constant, manipulating its values, or isolating the system with, e.g., a soundproof box or white-noise masking generator) or of *observation.* In some sciences (e.g., nuclear physics), it would be quite difficult to parse these theoretical claims from the theory being tested, but such is not the case in the behavioral sciences (cf. Meehl, 1983b, pp. 389-395). I treat a galvanometer used in studying galvanic skin response or a Skinner box as an instrument, and statements of general form that are relied on when such instruments are used in a psychological experiment as belonging to the set $A_i$. I am using the term narrowly, and it is sufficient for present purposes to stipulate that the theory of an instrument must not contain, explicitly or implicitly, any psychological constructs or theories. The electrochemical theory about an electrode on the skin belongs to $A_i$, but the "psychometric theory" of the Minnesota Multiphasic Personality Inventory (MMPI) or Rorschach belongs to $A_t$, not $A_i$. If we explain away a certain MMPI score in terms of the subject's non-cooperativeness or deficiency in English as shown by a high *F* score, such discourse belongs to psychology, although this may not be the branch of psychological theory we are interested in studying at the moment. The line between *T* and $A_t$ is somewhat fuzzy and, here again, is probably more so in physics and chemistry, where the instrumental processes themselves belong to the same theoretical domain as the theories under test, than in psychology. It is not necessary for what follows, and I do not wish to maintain that it is always possible, to make a clean distinction between *T* and $A_t$, but I have some suggestions to make along those lines.

7. In his discussion of the positive and negative heuristic, Lakatos (1970) lumped all the conjuncts on the left except T as part of the "protective belt," and maybe even portions of *T.* (Even *T* itself has a hard core and a periphery, which I discuss later.) Lakatos also subsumed both *disturbing particulars* (one way to violate $C_p$) and *incomplete statement of auxiliary general laws* ($A_t$), into his ceteris paribus clause. I think it is important to distinguish these, especially because, as Lakatos pointed out in his planetary examples, denying $C_p$ via conjecturing a new particular sometimes functions to turn an apparent falsifier into a corroborator. The discovery of Neptune as the origin of the apparent falsification of Kepler and Newton by the aberrant orbit of Uranus is a famous example from the history of science. Whereas when we deny $C_p$ by postulating an additional auxiliary *theory* $A_t$, this does not, at that point, function corroboratively but merely defensively, and gives rise to the problem of what kind of ad hockery we are engaged in, the good kind or the bad kind.

8. In the presence of what appears to be a falsifying protocol, the Lakatosian methodology prescribes a strategic retreat (a *Lakatosian defense,* I call it). *When* adoption of this strategy is warranted, instead of confessing immediately that *T* has been falsified and should be abandoned, remains to be discussed: In what follows immediately I consider the *literal truth of T*, because we can't discuss everything at once. In reality, a sensible psychologist would take it for granted that *T* itself is almost certainly imperfect, either in (a) the weak sense that it is *incomplete* or (b) the strong sense that it is, when taken literally, *false.* This involves the problem of verisimilitude, and the important Lakatosian distinction between saying that a theory is falsified and saying that one ought rationally to abandon it. In science, theories when falsified are not abandoned prior to a Kuhnian revolution (Kuhn, 1970), but are appraised as to their degree of verisimilitude, and attempts are made to patch them up. But in discussing the Lakatosian strategy of retreat, I initially set

aside the problem of verisimilitude of $T$ and reason as if we wish to defend it literally as it stands.

In our strategic retreat we may choose not to admit the falsifying protocol, a tactic that may include doubts regarding the instrumental auxiliaries $A_i$. Students are bothered by this tactic if they were taught a simplistic empiricism in undergraduate psychology classes and deem it sinful of an empiricist to stick to his theory and not "admit the facts." The thing to see here is that it is not a question of going against the facts, but of denying that an alleged fact is in reality a fact. What is available to the critical scholar is not the fact but some other scientist's sentence asserting it. As Lakatos emphasized, we have shining examples from the history of science of the success of this approach, as when the scientific community of physics did not admit Dayton C. Miller's protocol of an ether drift (which required a quarter of a century to explain as a thermal artifact), or Mendeleev's maintaining the correctness of his periodic table by insisting that the received atomic weights of gold and tellurium must be in error.

If we admit the falsifying protocol, accepting the instrumental auxiliary, we may then elect to challenge $C_p$. This is a plausible proceeding in psychology because we believe with near certainty that there are missing systematic factors. "Ceteris paribus" does not, of course, mean "all the factors not mentioned by us are equal for all subjects of the experiment." If that were the case, there would be no error term to go into the denominator of a significance test and no methodological prescriptions regarding stratified or random sampling. What the ceteris paribus clause says is that there are no *systematic* factors left unmentioned; as when, in path analysis, the individual differences in an output variable not attributable to endogenous variables in the path diagram are explained in terms of largely unnamed "disturbance factors" represented by an exogenous arrow $u$ whose influence, varying over individuals, is conjectured to be uncorrelated with the variables included in the diagram.

Suppose I am a psychopathologist studying motivation in schizophrenes, and I do so by exposing them to a social stimulus and seeing how this influences their perception of ambiguous objects in a tachistoscopic experiment. No psychologist supposes that we have a complete science of psycholinguistics assuring us that there could not be any cognitive nuisance factors influencing how our instructions are understood, factors that might be correlated with some of the patient characteristics that we include in our experimental design as "factors"; similarly, we do not assume that the theory of tachistoscopic perception is complete. Common sense tells us that both the importance and the dangerousness of $C_p$ are much greater in psychology than in chemistry or genetics. The ceteris paribus clause amounts to a very strong and highly improbable negative assertion, to wit, *nothing else is at work except factors that are totally random and therefore subject to being dealt with by our statistical methods.* For the ceteris paribus clause to be literally acceptable in most psychological research, one would have to make the absurd claim that whatever domain of theory is being studied (say, personality dynamics), *all other domains have been thoroughly researched,* and all the theoretical entities having causal efficacy on anything being manipulated or observed have been fully worked out! If that were the case, why are all those other psychologists still busy studying perception, learning, psycholinguistics, and so forth?

9. In conducting the strategic retreat in the presence of accepted falsifiers it is useful to think in terms of a theory as attempting to deal with several fact domains. One of the impresssive things about a science like physics is that it predicts and explains observations from domains that at the phenomenological level are nonoverlapping. It is again part of the received tradition of scientific "common sense" that a theory's ability to handle facts in qualitatively diverse domains is more impressive than its only handling a large number of particulars belonging to the same domain. Any working scientist is more impressed with 2 replications in each of 6 highly dissimilar experimental contexts than he is with 12 replications of the same experiment. Suppose $T$ is doing very well in several domains, and it has also succeeded with a few high-risk predictions in a subdomain in which also, however, the conjunction *$(T \,.\, A_t \,.\, C_p)$* has been clearly falsified. Then an obvious strategy is to amend the *domain $C_p$*. In physics, the same basic laws apply to everything we study. But in psychology one may reasonably conjecture that the trouble arises from the $C_p$ within the domain. For instance, suppose I study a psychodynamic problem in bipolar depressives by a structured inventory, a projective test, and a tachistoscopic experiment. My theory does well with the first two and does moderately well with the tachistoscopic setup, but also has several clear falsifications there. It is reasonable to wonder whether there is something about, say, the attention and information processing times of psychotically depressed patients that I haven't been considering, a special something that would not be expected to interfere with an untimed determinant of the Rorschach or in answering the verbal items of the MMPI. The psychologist has the option of moving around with some freedom in denying $C_p$ for a domain or a subdomain without getting into trouble in other theoretical derivations, and in this respect he is "safer" in challenging $C_p$ than the physicist or the astronomer.

10. A related situation exists with regard to the theoretical auxiliaries $A_t$ where one asks how widely $A_t$ is found in the various derivation chains in different domains before modifying it to deal with a subdomain falsification. A further criterion is the extent to which a certain auxiliary has been independently corroborated in other experiments not involving the $T$ of current interest. I am not aware of any rigorous treatment of this, and one may question whether such may be possible absent an empirical statistical study of the history of science. Stated qualitatively, the problem of adopting a strategy is simple: We want to preserve the derivation chains that have been doing well, so we don't want to challenge the ceteris paribus clause with the introduction of new theoretical entities or laws that we would then have no rational basis for denying efficacy in the other domains where the theory was doing well without them. We do not want to be guilty of gerrymandering the ad hockery we perform on our auxiliaries!

11. This strategic retreat—beginning with the reluctant admission of the falsifying protocol, then cooking up new auxiliaries by denial of the ceteris paribus clause in troublesome domains, and then challenging some of the former auxiliaries themselves—may finally result in recognizing that the program begins to look somewhat "degenerate," as Lakatos called it. If pursuing the positive heuristic leads to an excessive amount of ad hockery (any of Lakatos's, 1970, three kinds of ad hoc) the research program is called *degenerating.* If the adjustments made in the protective belt are

content increasing, empirically successful, and in some sense inspired by the leading ideas of the theory (rather than alien elements pasted on), the research program is said to be *progressive.* Feyerabend (1970) criticized this because one does not have an objective cutting score for how long an appearance of degeneration should continue before deciding to abandon the negative heuristic and challenge the hard core, but I do not find this persuasive. There can hardly be any such precise demarcation line, and given Feyerabend's general views it seems odd that he should demand one. The situation is the same as in many other pragmatic decision contexts. As more and more ad hockery piles up in the program, the psychological threshold (which will show individual differences from one scientist to another) for grave scepticism as to the hard core will be increasingly often passed, inducing an increasing number of able intellects to become suspicious about the hard core and to start thinking about a radically new theory. As pointed out in my 1967 article, one can easily find examples in soft psychology where the ad hockery is multifarious; but due to the flabby significance-test tradition, what is clearly a Lakatosian degenerating research program is viewed favorably simply because the successive stages of ad hockery suggested new experiments. The fact that the batting average of the predictions from the new experiments to test each ad hoc stage in the Lakatosian defense is poor will not bother a psychologist unfamiliar with the Popperian line.

12. Like the concept of verisimilitude, the metaconcept of core or central portions of a theory has not been given rigorous definition, and I am not able to offer one. It is obvious that some such distinction must, however loosely, be made. Intuitively one sees that in a particular theory some components are ubiquitous in dealing with the range of facts whereas others are not thus centrally located, although they are truly "part of the theory," as both the theorist and critics would usually agree. For example, if I describe myself as a "neo-Freudian" and you ask me why I qualify with the 'neo', I might say that I have doubts about the universality of the Oedipus complex, or that penis envy plays a crucial role in the psychopathology of women. This would not lead you to deny me the right to call myself a modified Freudian. In fact, Freud himself said, in his 1914 polemic on the history of the movement (see Freud, 1914/1957)—where we may assume he was at pains to be exact in demarcating what may be called 'psychoanalysis' and what does not deserve that appellation—that anyone who accepts the basic facts of transference and resistance may call himself a psychoanalyst whether he agrees with Freud in other respects or not. This is a remarkably broad definition. But if I told you that I was a modified Freudian who did not believe in the reality of unconscious mental processes, and I did not think that conflict played any appreciable role in the pathogenesis of neuroses, I would be talking nonsense. As another example, suppose I told you that I was a disciple of Skinner but that I had inserted a couple of special postulates about stimulus–stimulus (S–S) conditioning to deal with the nagging problem of latent learning, assuming that to have been satisfactorily replicated in the operant conditioning chamber. Skinner might not be entirely happy with this, but I would not be talking nonsense to describe myself as a modified Skinnerian. Whereas if I said I was a neo-Skinnerian, my amendment to Skinner's theoretical system being that reinforcement contingencies are of no special importance in understanding behavior, that would be nonsensical talk. These examples make it obvious that there is some kind of distinction between the hard core of the theory and its periphery.

At the risk of poaching on the logician's domain, I attempt to say something tentative about how this distinction might be usefully spelled out by those more competent. The main thing about the core concepts of a theory is that they recur when explaining facts in all (or almost all) of the phenomenal domains that the theory purports to address. We might formalize this "explanatory ubiquity" and try to define a *core postulate* as one that appears in every derivation chain. That doesn't quite work, because not every experiment involves *explicit* mention of a core postulate as so defined. Instead, there may be reference to a *concept* which is quantified and whose numerical value in a particular organism depends on past events whose mode of action is stated in the core postulate. For instance, in Hull's (1943) system, the law of acquisition of habit strength does not explicitly appear when we are studying the shape of the stimulus generalization gradient, which makes it look as if the habit strength postulate is not "core" to Hull's system in my ubiquitous sense. But, of course, the gradient has its peak at the point of conditioning, and it is because of that indirect reference that one might say that the habit strength postulate is core. If an experimenter presented us with a stimulus generalization curve apparently refuting Hull's theory, but omitted to tell us that the rats that determined particular points on his curve had been subjected to varying amounts of reinforcement with respect to the originally conditioned stimulus, that would be a gross piece of scientific malreporting.

So we might approach it instead by saying that if a certain *concept* appears in every derivation chain, either explicitly, or implicitly in that every derivation chain contains concepts that are theoretically defined by reference to it, that concept is a *core concept.* Then one might go on to say that a postulate of the theory consisting only of core concepts is a core postulate. As shown in the next section, I think a satisfactory explication of the concept of verisimilitude will depend on first formulating the core–peripheral distinction. That is, a theory that is qualitatively false in its core postulates has lower verisimilitude than one that is qualitatively correct in its core concepts or postulates but incorrect in several of its peripheral ones.

### Excursus: The Concept of Verisimilitude

It is unfortunate that the logician has not been able as yet to develop a rigorous explication of the verisimilitude concept ("truth-likeness"), because this concept is indispensable in metatheoretical discussion of theory appraisal. We cannot dispense with an important idea on the grounds that it has not been rigorously explicated, a proceeding that would be strange to follow in metatheoretical discourse when nobody insists on following it in the substantive discourse of a scientific theory proper. If we find we cannot get along without a fuzzy notion in our substantive theory, we make use of it and hope that sooner or later somebody will figure out how to expound it more rigorously. (On open concepts, see Cronbach & Meehl, 1955; Meehl, 1972, p. 21; Meehl, 1973b, p. 195; Meehl, 1986b, 1990b; Meehl & Golden, 1982; Pap, 1953, 1958, 1962). It is reasonable to adopt the same view toward metatheoretical concepts. The notion of *degrees of* *verisimilitude* does not conflict with the view that statements are either true or false, because a scientific theory doesn't consist of a single statement about "simples" (if there are any metaphysical simples!), but is a conjunction of interrelated statements about complexes. So, even in employing such a crude approach as a truth frequency count (which will not do as an explication of verisimilitude, although it has been tried), we recognize that some texts are more verisimilar than others. Not just a matter of philosophy of science, this obvious point is familiar to us from everyday life, history, journalism, courts of law, and so on. If a newspaper account describes an automobile accident and gets everything right except the middle initial of one of the participants, we say that it has very high verisimilitude. If it describes an accident that occurred, but gets one of the names wrong, as well as the numbering of the intersection, we think of it as a poor story but still containing some truth. If it's totally made up out of whole cloth by Dr. Goebbels, as the hoked up Polish attack on the Gleiwitz radio transmitter, we say it has zero verisimilitude. Similarly, in a court of law, impeachment of a witness by getting him to contradict himself does not lead a judge to instruct the jury to ignore every single statement that he made; instead they are supposed to assign some appropriate correction to the weight they give his testimony on the grounds of a clear inaccuracy in a certain respect. Up to now my discussion has spoken solely in terms of the *truth* of a theory and its auxiliaries. But, of course, every scientist in the back of his mind takes it for granted that even the best theory is likely to be an approximation to the true state of affairs. For this reason, *a falsification* of *T* does not necessarily result in an *abandonment* of *T,* in the sense of dropping it completely and starting from scratch with a new theory having no overlap in concepts or postulates with the one we abandoned. When the strategic retreat from the falsifying protocols, through the instrumental auxiliaries and statement of particular conditions, challenging the ceteris paribus clause in one or more fact domains, creating new auxiliaries and modifying old ones, has resulted in what appears to be a degenerating program but one not bad enough to give rise to a scientific revolution, what the scientist does is to begin looking for ways of amending *T* itself. This is a rational strategy to the extent that there are grounds for thinking that the theory, *although literally false,* possesses high verisimilitude. Verisimilitude is an ontological concept; that is, it refers to the relationship between the theory and the real world which the theory speaks about. *It is not* an epistemological concept; that is, it does not refer to the grounds of rational belief. I am going to adopt the working scientist's attitude in this matter, that verisimilitude is correlated, in the long run, with evidentiary support, again relying on future philosophers of science to show why this relationship might be expected to obtain (but cf. Meehl, 1990a). Keeping the distinction in mind, we postulate a stochastic connection between the degree of evidentiary support, the number, variety, and stringency of empirical tests that the theory has passed or failed, and its verisimilitude, its closeness to objective reality.

Efforts to define verisimilitude as Popper first did, by some kind of relation between truth and falsity content, got into a variety of difficulties, including technical problems of measure theory and the like. It seems generally agreed that these approaches will not wash (cf. references in Brink & Heidema, 1987; Goldstick & O'Neill, 1988). I think that metatheory should go at it in a somewhat different way along the following lines, which I do not claim to be a rigorous explication. Suppose we have a theory $T_1$ and another theory $T_2$ and we ask how similar they are to one another. It seems to me that the first thing a working scientist asks when examining theories is what kinds of entities they speak of. So far as I can tell, there are only a half dozen different *kinds* of constructs found in any of the sciences, namely (a) substances, (b) structures, (c) events, (d) states, (e) dispositions, and (f) fields. The first thing (see Figure 1) we do in comparing $T_1$ and $T_2$ is to inquire whether they postulate similar *lists* of theoretical constructs .As a clear, crude case, if $T_1$ and $T_2$ each conjecture the same kinds of constructs (e.g., one substance and two structures) and propose that the substances and structures have such-and-such dispositions (equal in number), we would suspect that however different the terminology or associated imagery of the theorists, their theories were quite similar, perhaps identical in semantic content. Next we ask how these theoretical entities are related to one another. For example, structures of one kind have causal relations to structures of another kind that then jointly combine to bring about such-and-such a state in a substance. In the network metaphor (Cronbach & Meehl, 1955), if we could superimpose the two nets on each other so that entities that constitute the nodes of the net are connected by causal or compositional laws in the same ways in $T_1$ and $T_2$, then we would consider them isomorphic. The functional dynamic laws connecting events or states of the various theoretical entities can be specified in varying degrees of mathematical detail (cf. MacCorquodale & Meehl, 1954, pp. 214-215). Weakly, one may assert merely that when entity $E_1$ undergoes an increment in its state $S_1$, then entity $E_2$ undergoes an increment in its state $S_2$. Here we know only that $dx_2/dx_1 > 0$ in both theories. Stronger is a comparative claim about two causal influences, that $\delta y/\delta x > \delta y/\delta z$ everywhere. Or we may be prepared to conjecture that $d^2y/\ dx^2 < 0$ everywhere (i.e., the functional dependence of $y$ on $x$ is decelerated). Increasing detail involves comparison of mixed partial derivatives, then specification of function form (hyperbola? log? growth function?), and, finally, assigning quantitative values to the parameters. For the most part, these specifications are lexically ordered, in Rawls's (1971) sense. It wouldn't make sense to compare the parameters of a hyperbola in $T_1$ with those of a growth function in $T_2$. So we don't reach that question unless the function forms are the same in $T_1$ and $T_2$. Nor could we ask whether the function forms relating states, events, or dispositions in two theoretical entities were the same if in one theory these entities have a strand in the nomological network connecting the two nodes and in the other they are not connected, so that if they are correlated, their correlation is not due to the operation of Aristotle's "efficient causality" between them. Obviously, none of these formal questions would make any sense if the theories differed completely as to the kinds of entities they postulated to exist.

I suggest that this kind of approach is closer to the way scientists actually think than logicians' infinite consequence-class of possible falsifiers and the like, and that it would not run into the mathematical and logical paradoxes that the logicians' approach gives rise to. I do not think it absurd to imagine some sort of crude quantitative index of the similarity of two theories that could be constructed on the basis of the theoretical properties I have listed, but that is music of the future. Suppose we did have

**Theory Specifications**

**I.** Type of entity postulated (substance, structure, event, state, disposition, field)

**II.** Compositional, developmental, or efficient-causal connections between the entities in (**I**)

**III.** Signs of first and second derivatives of functional dynamic laws in (**II**)

**IV.** Ordering relationships among the derivatives in (**II**)

**V.** Signs of mixed second order partial derivatives ("interactions") in (**II**)

**VI.** Function forms (e.g., linear? logarithmic? exponential?) in (**II**)

**VII.** Trans-situationality of parameters in (**VI**)

**VIII.** Quantitative relations among parameters in (**VI**)

**IX.** Numerical values of parameters in (**VI**)

**Figure 1. Progressively stronger specifications in comparing two theories (similitude).**

some such way of expressing how similar two theories $T_i$ and $Tj$ are to each other. Now consider theory $T_{OJ}$, the theory my former philosophy colleague Wilfred Sellars used to call "Omniscient Jones's" theory—that is, the true theory of the domain. Then the similarity of $T_i$ to $T_{OJ}$ defines the verisimilitude of $T_i$.

### Two Principles That Warrant Lakatosian Defense of a Theory

The reader will have noticed that up to this point I have said almost nothing about significance tests, or about statistics generally. Although a theory's merit is a matter of degree rather than a yes-or-no question (as it is treated in null hypothesis refutation and in some but not all of Popper), I do not think "what degree of merit" is best expressed in significance-test terms, or even by specifying a confidence belt. In spelling out how to conceive and implement Serlin and Lapsley's (1985) "good enough" principle, my emphasis remains different from theirs, although my present position is not the strong Popperian falsification one that they criticized, as I now agree with them that falsification is not the crux, because we know the theory is imperfect.

All psychological theories are imperfect (defective), at least in the sense of being *incomplete.* Most of them are, in addition, *false* as far as they go, almost certainly false when they go to the point of stating a mathematical law. I formerly made the mistake of saying that all scientific theories are false, holding that they are all lies, so the question is how can we tell the theories that are white lies from those that are black lies, and how do we move the gray lies in the white-lie direction? (See, in this connection, Cartwright, 1983.) This is not usually correct except for (a) quantitative theories or (b) cosmological theories, as Feyerabend calls them, theories that say something about everything there is. Cartwright, in her fascinating book, admitted to having made that mistake concerning laws until a colleague pointed out to her that nonquantitative theories in several domains of science (e.g., biology) can be literally true (Cartwright, 1983, pp. 46, 54-55). Even quantitative theories can be made literally true by putting bounds on the numbers instead of giving point values. What happened historically was surprise at finding the paradigm of all scientific theories, which everybody tried to emulate, namely Newton's, to be literally false. It was natural to think that if this great paradigm and paragon of scientific theorizing could turn out after a couple of successful centuries to be false, then probably all theories are false, if "only a little bit" so. But Newton's theory took the grave risks of (a) being cosmological, and (b) stating strict quantitative laws, and therefore ultimately was falsified. If we consider, say, Crick and Watson's theory of the gene, does anybody seriously think that will ever be falsified? Stated in qualitative terms, does anybody think that science will ever find that they were wrong in conjecturing that genes are composed of triplets of codons, arranged with a helix whose frame is provided by deoxyribose and the phosphate radical? Does anyone conceive that future research could show that the sun is not, after all, a big ball of hot gas—mostly hydrogen—but that it is a glowing gigantic iron cannonball (as Anaxagoras conjectured), or Apollo's chariot? We may yet learn that the human liver has some functions presently unknown. But surely no one thinks that future physiology may conclude that, contrary to what we believe today, the liver does not store glycogen, or secrete bile, or detoxify.

So it is incorrect to say that all theories are false. It depends on what kinds of theories, and how they are stated. In psychology, they are at least all defective, in the sense of being incomplete. This obvious metatheoretical truth gives rise to an interesting point concerning aspects of verisimilitude, the relation between "the whole truth" (incomplete) and "nothing but the truth" (literally false). When an incomplete theory is used in a derivation chain to predict the results of an experimental or statistical study, the derivation does not go through rigorously absent the ceteris paribus clause $C_p$, almost always false in psychology. So that whereas $T$ may not, so far as it goes, make literally false statements about the way things are, whenever $T$ is employed to explain or predict facts, the derivation chain utilized, without which $T$ would not be an empirically testable theory, is always literally false, because the theory's incompleteness, or our failure to know certain additional auxiliaries $A_1, A_2, \ldots, A_m$, falsifies $C_p$.

As a general statement about the Serlin–Lapsley principle, I assert that because, in psychology, we *know* that the verisimilitude is imperfect, we do not want to equate "good enough" with "close enough numerically to continue believing it true." Rather we want to equate "good enough" with some such notion as "having enough verisimilitude to warrant continued effort at testing it, amending it, and fiddling in honest ad hockery (not ad hoc of Lakatos's three forbidden kinds) with the auxiliaries." I would propose two subprinciples that I think suffice, when conjoined, to explicate Serlin and Lapsley's principle on this general basis. The first one might be called the "track record" or "money in the bank" principle. Because it gives conditions under which it is rational to conduct a Lakatosian defense ("strategic retreat" from the protocol back to the theory's hard core), one could label it the Lakatos principle, and I do so. The second is the "damn strange coincidence" criterion, which I label Salmon's principle for Wesley Salmon (1984), who coined the phrase and made the argument explicitly. Lakatos's principle says that we are warranted in continuing to conjecture that a theory has high verisimilitude when it has accumulated "money in the bank" by passing several stiff tests. If it has not done this, for instance, if the tests consist of mere refutations of the null hypothesis, the majority of which have panned out but a minority not, it is not rational to adopt the Lakatosian heuristic and engage in strategic defensive retreat, because we had feeble grounds for favorably appraising the theory as it stood *before* it began to run into the apparent falsifiers. Without some niceties found in his incisive and powerful exposition, important to philosophers but not to us here, I formulate my version of the Lakatos principle thus: Accepting the neo-Popperian view that it is inadvisable to persist in defending a theory against apparent falsifications by ad hoc adjustments (three kinds), *the rationale for defending by non-ad hoc adjustments lies in the theory having accumulated credit by strong successes, having lots of money in the bank.* Although persistence against this advice has been known sometimes to succeed, one should do it rarely, knowingly, and with explicit public recognition that either the theory never had much money in the bank, or that even though it has had good credit, the defensive research program is now degenerating.

Anticipating a critic's objection that Lakatos has not explicitly stated this, I am not aiming here to provide a history of science exegesis of his writings; rather I am formulating, especially for psychologists, the "Big Lesson" he has to teach *us,* honoring the man eponymically in passing. Imre had a complex and subtle mind, as shown, for instance, by the rich proliferation of footnotes in his writings, none of them superfluous. (It would be remarkable if all those intellectual sparks were entirely consistent!) I am aware that he countenanced rare deviations from his "antidegeneration" principles, as in the following response to objections by Feyerabend and Musgrave:

> Let me try to explain why such objections are beside the point. One may rationally stick to a degenerating programme until it is overtaken by a rival *and even after.* What one must *not* do is to deny its poor public record. Both Feyerabend and Kuhn conflate *methodological* appraisal of a programme with firm *heuristic* advice about what to do. ...It is perfectly rational to play a risky game: what is irrational is to deceive oneself about the risk. (Lakatos, 1971, p. 117)

One supposes the "rationality" of this (normally contraindicated) stance would lie in the individual scientist's values, lifestyle, self-confidence, even "personal track-record" as a strangely successful maverick who has taken seemingly foolish cognitive gambles and won. It is a social fact that some scientists have sounder intuitions than others, and those who sense that about themselves may rationally choose to march to a different drum. But note the somewhat shocking paragraph that follows this concessive, "tolerant" text:

> This does not mean as much licence as might appear for those who stick to a degenerating programme. For they can do this mostly only in private. Editors of scientific journals should refuse to publish their papers which will, in general, contain either solemn reassertions of their position or absorption of counterevidence (or even of rival programmes) by *ad hoc,* linguistic adjustments. Research foundations, too, should refuse money. (Lakatos, 1971, p. 117)

So I think it legitimate to christen with his name my short formulation of what is clearly the main thrust of his neo-Popperian position.

The way a theory accumulates sizable amounts in the bank is by making risky predictions. But unlike unmodified Popper, we are not looking on those risky predictions primarily as ways of deciding whether the theory is literally false. Rather we suspect it would not have passed *some* risky tests, and done reasonably well (come numerically close) in others, if it lacked verisimilitude. My criticism of the conventional significance testing procedure still stands, despite Serlin and Lapsley, because it does not involve a series of "damn strange coincidences." Salmon's principle I formulate thus: *The main way a theory gets money in the bank is by predicting facts that, absent the theory, would be antecedently improbable.* When predictions are quantitative, "near misses" count favorably along with "clear hits," both being unlikely coincidences. Conventional significance testing plays a minor and misleading role in implementing either of these two principles. Even confidence belts, although more respectable and more in harmony with the practice of advanced sciences, play a lesser role than I formerly supposed.

In this connection I note that the physicist's, chemist's, and astronomer's near equivalent of what we call a "significance test" is the attachment of a standard error to a set of observations. Sometimes this has the function of telling us how trustworthy an estimate is of a parameter (working within a theory that is not considered problematic). But sometimes it has the different function of testing whether the distribution of observations is compatible with the predictions of a substantive theory. As I pointed out in my 1967 article, when the physicist uses a probable error in this second way, improvement in the quality and number of measurements leading to a lessened standard error subjects the theory to a greater risk of falsification, because here a "significant deviation" means *a deviation from the predicted point value or curve type.* That is how Karl Pearson's original invention of chi square at the turn of the century worked. His idea of chi square was as an indicator of frequency *discordance,* asking for example, does an observed distribution depart significantly from the frequencies in class intervals as given by the Gaussian (or other theoretical) function? This I call the *strong use* of a significance test. But then occurs a development in the use of chi square, at Pearson's own hands admittedly, in which the "theoretical" or "expected" values of cell frequencies, rather than being positively generated by an affirmative substantive theory generating a certain mathematical form, are instead specified by the hypothesis that two variables are *not* related to one another. So the expected values of cell tallies are provided by multiplying the marginals on the hypothesis of independence, using the product theorem of the probability calculus. There is, of course, nothing wrong with the mathematics of that procedure. But social scientists seem unaware of the great shift methodologically that takes place in that reverse-direction use of a significance test, where now the substantive theory is supported by the achievement of significance in departing from the "empty" hypothesis that two things are unrelated. In the strong use of a significance test, the more precise the experiment, the more dangerous for the theory. Whereas the social scientist's use of chi square in a fourfold table, where $H_0$ is that "These things are not related," I call the *weak use.* Here, getting a significant result depends solely on the statistical power function, because the null hypothesis is always literally false.

In what follows it is important to keep in mind the fundamental distinction between a substantive theory $T$ and a statistical hypothesis $H$. Textbooks and lecturers on statistics do not stress the distinction, and some do not even *mention* it by so much as a single monitory sentence. This grave pedagogical omission results in the tendency of students to conflate refuting $H_0$ with proving the counternull, $-H_0$, which then is immediately identified in their minds with "proving $T$." This tempting line of thought thus combines a mistake in the strictly statistical reasoning with a further mistake in logical reasoning, affirming the consequent in empirical inference. In sciences where individuals differ, for known or unknown reasons, and even in sciences where individual differences play no role but measurements are subject to error, the observed numerical values, whether of degree (metric) or of frequency (count, rate), are subject to fluctuation, so we call in the statistician to help us with that part of the problem. If there were a science having infallible measuring instruments and in which the individuals studied showed no individual differences, so that neither measuring error nor sampling error was a relevant concept, then conventional statistics would be a minor branch of mathematics of little scientific relevance. But that glorious state of *observational* affairs would do nothing to ameliorate the problems of inductive logic, Theoretical inferences are always ampliative and do not flow as a deductive consequence of any finite class of observation statements. The purely logical point here is, as I said earlier, that empirical inference from fact to theory is in an invalid figure of the implicative syllogism, so *formally* the theorist's transition is the fallacy of affirming the consequent (hence, Morris Raphael Cohen's malicious witticism). Speaking methodologically, this formal point corresponds to saying, "... but there could be other theories that would explain the facts equally well." The poor social scientist, confronted with the twofold problem of dangerous inferential passage (right-to-left) in Figure 2 is rescued as to the $(H \rightarrow O)$ problem by the statistician. Comforted by these "objective" inferential tools (formulas and tables), the social scientist easily forgets about the far more serious, and less tractable, $(T \rightarrow H)$ problem, which the statistics text does not address.

One reason why psychologists in the soft areas naively think that they have strongly proved a weak theory by a few significant chi squares on fourfold tables is that in their education they learned to conflate *statistical significance* with the broader concept of *evidentiary support.* So they are tempted to believe that if there is nothing wrong with the experimental design, or in the choice of statistic used to test significance, they are "safe" in

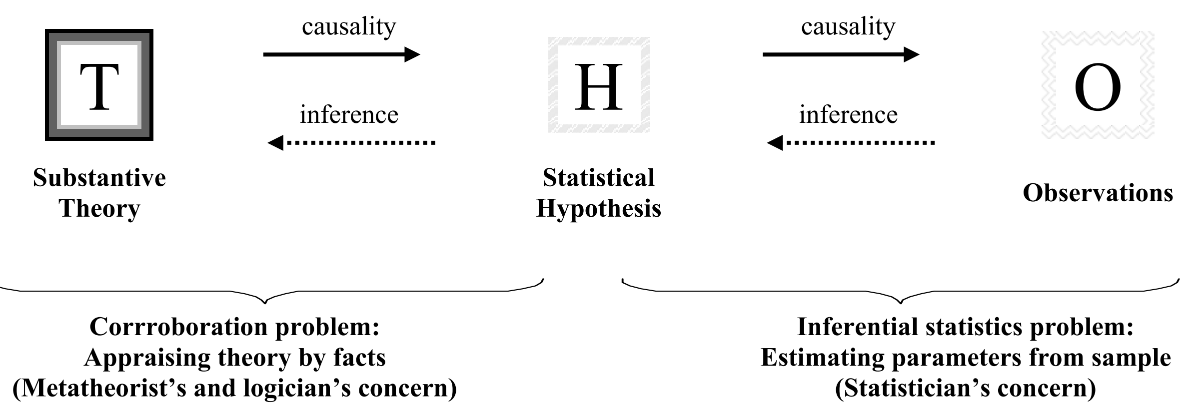


**Figure 2. Causal and inferential relations between substantive theory, statistical hypothesis, and observational data.**

concluding for the verisimilitude of a theory. Pedagogically, I have found the quickest way to dispel that comforting illusion is to put the question, "Assume you *had* the parameter; what would you know, and how confidently?"

If the way in which a substantive theory gets money in the bank (thereby warranting us rationally to engage in strategic retreat rather than to abandon it forthwith) is by satisfying Salmon's principle, we must now examine how that works. Successful prediction of numerical point values is the easiest one to explain, although as I have pointed out elsewhere (Meehl, 1978) there are other pretty good ones, such as predicting function forms and rank orders. I suppose that underlying Salmon's "damn strange coincidence" notion is a basic maxim expressing scientific optimism (or "animal faith" metaphysics), something like this: "If your aim is causal understanding of the world, do not adopt a policy of attributing replicable orderliness of observations to a damn strange coincidence." Salmon's favorite example (also my favorite in teaching this material to psychologists) is the convergence of numerical values for Avogadro's number $N$ by 13 qualitatively disparate avenues of evidence, as set forth by Nobel laureate Perrin in his classic work *Atoms* (1913/1916; see also Nye, 1972, or the excellent shorter treatment by Salmon, 1984). Up to that time many physicists, including such distinguished ones as Mach, Ostwald, Duhem, Le Chatelier, and Poincaré, denied the real existence of molecules, considering them merely as a useful computational device, a kind of handy "scientific fiction." In his book, Perrin pulled together 13 different ways of estimating the number of molecules in a mole, ranging from the fact that the sky is blue to the distribution of displacements of a Brownian particle, the mathematics of this having been derived by Einstein in 1905. These qualitatively disparate observational avenues for estimating the number of conjectured small particles in a gram molecular weight of a substance all came out with values approximately $6 \times 10^{23}$.

This famous physical-science example highlights the differences among (a) the weak use of significance tests to provide feeble "confirmation" of weak theories, (b) the strong use of significance tests in discorroborating strong theories, and (c) the third approach—which I advocate—that is more characteristic of the developed sciences, bypassing the statistical significance problem (except for special purposes like estimating constants within an already corroborated theory), namely, that of corroborating strong theories by Salmon's principle. It is easier to explain examples from Salmon's book than from the 13 relied on by Perrin, so I use three of his. One way of estimating Avogadro's number is via alpha decay. Because alpha particles are helium nuclei, and the number given off by a radioactive substance per time unit can be accurately measured by scintillation technique, and because alpha particles pick up electrons to become helium atoms, one can estimate the number of helium atoms produced in a container after alpha decay by counting scintillations. Then one simply weights the resultant quantity of helium to calculate molecules per mole. Second, starting with the conjecture that X-rays are very short light waves (beyond ultraviolet) plus the conjecture of the molecular theory of matter, considering the wave lengths of the X-rays and the diffraction produced when they pass through a crystal, one can estimate the spacing between atoms in the crystal and, via that, Avogadro's number. Third, from electrochemistry, knowing that it takes a charge of one electron to deposit an ion at the cathode of a silver chloride solution, on the basis of knowing the number of coulombs required to deposit one mole of silver, one can estimate Avogadro's number.

Suppose the theory were too weak to predict anything but monotone relationships between these variables. Suppose the theory merely said that you should get more helium from capturing alpha particles in a glass tube if you wait longer, that the distances between diffraction lines should be different between "hard" and "soft" X-rays, and that you should get more silver deposited at the cathode when a strong current passes through the electrolyte than when the current is a weak one. This would give us three directional predictions, and speaking nonparametrically, one might say that if they all panned out (as of course they would if it had been done this way) the probability that all three would come out in the right direction would be $p = .125$. This is marginal "significance." More to the point, suppose that at that level of significance we accept the statement that all three of these monotone relationships hold. This "$x$ is greater than $y$" finding, despite being in three qualitatively distinct domains, would hardly have convinced molecular unbelievers like Ostwald, whereas he threw in the sponge within a year of Perrin's 1908 paper (eight methods). We see here that there is a second big inferential step, *after* having concluded that the observations are not a matter of "chance." This is simply because we know that many theories, including continuous fluid theories and goodness knows what others, would be equally able to derive the algebraic sign of our results, without assuming the existence of molecules. In the electrolytic example, if we don't turn on the current, no silver is deposited. In a minute's flow, we get a tiny amount. We say "more yields more," that is, $dy/dx > 0$ throughout. Obviously, this observational result, which would be deducible from many different theories, does not strongly corroborate the specific molecular theory, merely one among all theories that would yield a monotone increasing function, relating amount to time. We know, even if we haven't yet worked hard at it, that the human mind is ingenious, and many clever scientists, if they set their minds to it, could concoct a variety of plausible nonmolecular theories that would explain more silver being deposited if the current flows longer.

Consider next the strong use of significance tests, going in the opposite direction, in which reaching statistical significance constitutes a falsifier of the substantive theory. The $F$ test did not exist in Perrin's day, although something similar to it, the Lexis ratio, did. But neither he nor anybody else bothered to ask whether the 13 values of Avogadro's number obtained by these quailtatively diverse avenues "differed significantly" from one another. I don't know if a contemporary Fisherian would fault them for not doing this, but I certainly would not. There is, of course, a special problem that arises here because the number being estimated is a theoretical quantity, and it differs numerically from the observational value not mainly because of sampling error—which is what conventional social science statistics always focus on, I think mistakenly—but because there is a chain of probabilistic inference running from the *qualitative* statements interpreting the formalism of the theory, to the observations. That is why a Fisherian complaint that you shouldn't need 13 statistical estimators of the same quantity if they're good estimators (meaning that they are maximum likelihood estimators, or *MLE*s) because, if they are, they will be both sufficient and efficient, is senseless in this context. An objection about sufficiency would totally miss the point. It conflates the mathematical question of estimating a parameter by random sampling from a specified physical distribution of measures, with the completely different (epistemic, not mathematical) point about converging lines of evidence. Perrin's reasoning cannot plausibly be represented along Fisherian lines. The qualitative diversity of the data base, permitting inference to an unobserved theoretical entity, is *not at all* the same kind of question as whether I have used an *MLE* of the variance of soldiers drawn as a random sample from the regiment.

Bypassing those niceties, let us imagine that, despite the fact that it's an inference via a conjectural theoretical chain of causes, we agree to treat the "distribution" of numbers (estimating Avogadro's constant in the 13 different ways) as a Fisherian statistical matter. We do an $F$ test to see whether they "differ significantly," which is a function of random measurement errors but also, and more important, of the *systematic errors* due to experimental bias arising from the unavoidable idealizations, especially the theoretical auxiliaries. Neither Perrin nor anybody else thought that those derivations were free of idealizations and approximations. Three sources of error exist that are not random and, hence, not taken care of by probability theory. First, the theoretical concepts are idealized in the interpretive text. Second, the formalism is approximative (e.g., terms in a Taylor expansion of an unknown function are dropped). Third, physical constants of viscosity, density, charge, and so forth are relied on without proof that their estimates are unbiased. So we may take it for granted, especially because a large number of measurements were made by each method, that the degrees of freedom above and below would give us a significant $F$ test. If we take a simplistic view of the kind Lakatos (1968, 1970) called $\text{Popper}_0$ (I agree with Popper that no such person exists), we would say that the strong use of the $F$ test has falsified the molecular theory.

Now no sensible physicist would have said that, nor should they have. Why not? Because we knew, before we started, that the theory had imperfect verisimilitude, and that some of the numerical values involved in those auxiliaries were inaccurate. So even this strong use of significance testing of the kind that occurs for certain purposes in the developed sciences would be an abuse if it were taken to mean not only falsification but abandonment. In this instance it doesn't even falsify the molecular theory, because of the problematic and approximative auxiliaries.

If significance testing had been applied by Perrin, a weak test of the social science type would give the "right answer" in confirming the molecular theory, but would confirm it only very weakly, and would not have convinced the fictionist skeptics. The strong use would have correctly falsified the theory-cum-auxiliary conjunction on the left of our Popperian equation, showing something we already knew before we did the experiments, namely, taken literally as it stands, the theory, together with the auxiliaries, is false. The first use gives us a correct answer, feebly supported. The second use gives us a correct answer we already know, and if the second one taken as a falsifier were translated into theory abandonment (which Lakatos, making a throat-cutting motion, called "instant rationality") we would be making a tragic scientific mistake.

What happened here, historically, without either such weak or strong significance testing? What happened in the history of science is what ought to have happened in a rational reconstruction; namely, physicists realized that if there were not any such things as molecules, then a set of 13 experimental procedures whose whole rationale is based on counting them could not have given such convergent numerical results except by a "damn strange coincidence." Following Salmon's principle, they decided not to treat it as a damn strange coincidence, but took it to be a strong corroboration for the existence of the theoretical entities that the 13 methods set out to count. If there aren't any molecules, derivation chains from 13 qualitatively diverse data domains whose whole rationale in the interpretive text, and the justification for steps in the mathematics, are based on the notion that the experiment is counting them, should not give the same answer. Simply put (as Poincaré said in his recantation), if 13 different ways to count molecules yield the same number, then there must be something being counted! And the point is *not* whether these 13 answers were "significantly different" from one another, which they doubtless were. The point is that *they were all of the same order of magnitude,* namely, $10^{23}$. (Psychologists are in the habit of using the phrase "order of magnitude" to mean "about the same," which is a sloppy use; it should be replaced by the physicist's and engineer's use, which is the exponent on base 10.)

You may say that this last is a probabilistic argument, whether one chooses to numerify it or not. Surely there is some sense in which this is rather like a significance test? I suppose there is. But I don't know how much it helps to formalize it to give a numerical value. One can do so, provided one is willing to make use of the old "principle of indifference" linked to the Leibnizian "principle of sufficient reason." One might here instead speak, as some Bayesians have, of the "principle of insufficient reason." One may divide a range of conceivable values into equal intervals and ask what is the probability, by chance, of falling into one of them? This was the basis of the classical Laplacian definition of the probability concept by the notion of "equally likely ways." This definition became unpopular (a) because of an alleged circularity in the notion of "equally likely" as a way of defining the concept "probability," (b) because of the paradoxes of geometrical probability, and (c) because of abuses of the principle of indifference, when combined with Bayes's theorem, to generate unacceptable consequences, such as Laplace's famous computation of the probability that the sun will rise tomorrow if we know how many times it has risen in the past. The deathblow to overdoing this kind of a priori range business was given by Fisher (1925, 1937) in the introductory chapter of his first book. Nevertheless, logicians (and some statisticians) have found it unavoidable, *under certain circumstances,* to think along those lines, and in recent years the ascendancy of Bayesian statisticians and philosophers of science has again made the idea of slicing up the range into equal intervals a priori a respectable move. I gather that the consensus among statisticians and logicians today is that it is respectable, problematic, or sinful depending on the context; and I suggest that Perrin's situation is one of those that makes it an acceptable kind of reasoning. If we wanted to generate a number to satisfy persons who don't like the notion of probability except as an expected relative frequency, we could proceed as follows. We could say that some familiar common-sense considerations about compressibility, the smallest things we can see with the microscope, and the like, entitle us to say that if there are any molecules, there can't conceivably be less than $10^3$ per mole. We don't know what the upper a priori limit is, so to be conservative we set the upper limit at the observed value, saying that the a priori possibilities for Avogadro's number do not go past order of magnitude $10^{23}$. Now suppose that there aren't any molecules, or anything like molecules, to be counted. Then all these derivation chains go through a mess of formalism that is empirically meaningless, not only in the sense that there is no interpretive text that gives meaning to the variables of the formalism, but in most of the derivation chains (I suspect all of them if you look closely) the mathematics itself doesn't go through without the embedding text. So all these derivations amount to a heap of nothing. If we agree to divide the numerical range from $10^4$ to $10^{23}$ into 20 subintervals (I leave it to the Bayesians to decide whether we should treat them this way or take logarithms; it doesn't matter here) then one may ask what is the probability, because the whole thing is sheer nonsense, that we would get three values in the same interval? If the theory makes the numerical prediction of approximately $6 \times 10^{23}$, the prediction is that all three will fall in the top interval, and the probability of getting that right "by chance" is $20^{-3}$. If the theory were too weak to give us the numerical value, but merely said that the *same* value should be reached by the three empirical avenues, then we could take one as the reference value, and the probability of the other two falling in the same interval as the chosen one would now be $20^{-2}$ ($p = .0025$). So for Perrin's table of 13 to agree (order of magnitude) "by chance" has minuscule odds, over a quadrillion-to-one against.

We contrast a theory sufficiently strong to generate a numerical point prediction with one too weak to do that, but strong enough to deduce that an unspecified numerical value characterizing a theoretical entity *should be the same when arrived at by two or more different observational avenues.* Such a distinction has a special importance in the behavioral sciences, because we are almost never in a position to do the first, but sometimes (how often?) we can do the second. The Perrin example shows that when "background knowledge," as the Bayesians call it, permits us to set up a rough range of a priori possibilities for an unknown numerical value, corroboration of a theory of only moderate strength can go up exponentially with the number of observational avenues by virtue of numerical agreement between two or more inferred values, despite none of them singly being theoretically deducible.

In psychopathology, for example, one is often interested in the question whether a certain nosological entity is taxonic, a true type, species, or "disease entity," or is merely a group of patients lying in an extreme region of the descriptor hyper-space. The conjecture that a taxon exists generates theorems that provide what I have called *consistency tests* for a latent taxonic model, but usually our theory will not be sufficient to specify the base rate of the conjectured latent taxon. So satisfaction of these consistency tests within allowable tolerances corroborates the taxonic conjecture, and permits an estimate of the taxon base rate, despite the fact that the theory would not have enabled us to derive that rate beforehand (Meehl, 1973a; Meehl & Golden, 1982).

Another example involves estimating the completeness of the fossil record, defined theoretically as what proportion of the species of some category (e.g., order Carnivora) have been found at least once as a fossil, so we know of the existence of that extinct species. Evolutionary theory does not enable us to make an estimate of that completeness index, but it should be possible to estimate the completeness index by multiple methods (Meehl, 1983a). If one asks whether such consistency tests are intended to validate the methods or, *assuming* the validity of the statistical methods, to raise our confidence in the numerical value of the index, that question is wrongly put, because the methodological situation is that we do both at once.

As pointed out in the cited article (Meehl, 1983a), a nice example of this from the history of physics was the crystallographic prediction of X-ray diffraction patterns on the conjecture that X-rays were electromagnetic radiation shorter than the ultraviolet and that crystals were atoms arranged in lattices that functioned in the same way with respect to X-rays as humanly made diffraction gratings function with respect to visible light. There is no basis on which the philosopher of science could decide at that stage in the history of physics whether the molecular theory of matter, and specifically the lattice conception of a crystal was an auxiliary, with the conjecture as to the nature of X-rays being the main theory under test, or the other way around. Derivation of the quantitative law went through, given the *conjunction* of these two theoretical conjectures and for the results to have panned out if either conjecture were false would have been a Salmonian coincidence. A physicist who accepted the molecular theory of matter but was doubtful as to the nature of X-rays, and another who looked at it the other way around, would have interchanged what each saw as the main conjecture of interest and the auxiliary, but logically at that stage of knowledge no such clear distinction could be drawn.

Another nice instance is the Van der Waals correction in the Boyle–Charles gas law where a prima facie falsifier—namely, that the derived gas law $PV = RT$ breaks down under extremes of density and pressure—is turned into a corroborator of the amended theory. The original derivation falsely conjectured as an idealization (which the theorists knew to be false taken literally) that the molecules in the gas occupy no space and have no attractive forces between them. Van der Waals made a subtraction from the observed volume term for the volume occupied by the molecules, and added to the observed pressure a term based on the notion that the mutual attraction of molecules weeds out a few of the slow ones in collisions just before they hit the wall. Because it takes two to make a collision, and the chances of a collision and hence the frequency vary as the squared density, which is the reciprocal of the square of the volume, his correction term is some constant divided by the square of the volume. But the point is that neither the value of this constant, nor of the constant referring to the space that molecules occupy, was theoretically derivable. These constants have to be found by a curve-fitting process, but the important point is that the curve, which now becomes somewhat complicated, $(P + a/V^2)(V - b) = RT$, does much better; and for the data to fit that function as well as they do would be a damn strange coincidence if there weren't any molecules acting the way the kinetic theory conjectures them to act.

Social scientists should not assume that the more developed sciences always have theories capable of generating numerical point values because that is historically not true. Far instance, Wien's law, derived in 1893, dealing with the spectral distribution of blackbody radiation, stated that for various temperatures of the blackbody, the energy density associated with a certain wavelength would be "some function" of the product of the wavelength and the Kelvin temperature, divided by the fifth power of the wavelength. The theory was too weak to say what that function was, but when one graphs the data points for several widely separated Kelvin temperatures, one gets a smooth curve with all the temperatures falling neatly on it (Eisberg, 1961, p. 50).

I venture to suggest that we psychologists have been less ingenious and resourceful than we might have been in working along these consistency-test lines because of a strange combination of optimism and pessimism. The optimism derives from uncritical acceptance of significance testing, almost always in its weak form, not realizing that this is a feeble way of appraising theories. The pessimism is because we cannot imagine, especially in the soft areas, concocting theories strong enough to generate numerical point predictions. It is important to see that intermediate strengths exist, where the theory is only moderately strong but is at least capable of deriving observational consequences about numerical agreements via qualitatively diverse observational avenues. I have made some constructive suggestions about this elsewhere (Meehl, 1990c), the most important of which is that the training of psychologists (even in the soft areas) should include a good deal more mathematics than is presently the case. I mean *mathematics, not statistics.*

All this is fairly straightforward contemporary philosophy of science. Now we come to one of those notions which, like verisimilitude, is crucial and unavoidable, but which cannot be rigorously explicated at the present time. What is it that makes a successful theory-mediated prediction (whether of a numerical value, or that, within tolerance, there should be good agreement between two or more numerical values none of which is theoretically predictable, but that the structural model says should agree when arrived at via different avenues) a sufficiently strange coincidence (absent the theory) that it gives high corroboration to the theory? The appropriate mental set in considering this question is different from the one that psychologists acquire from their exposure to courses in statistics, where the emphasis is on the deviation of a sample statistic from a population parameter. Whether one expresses this kind of "accuracy" as a standard error in physical units, or as a pure number the way engineers frequently do (percentage of the observed or inferred true value), neither of these gets at the main point of theory corroboration via successful numerical predictions. A standard error that is small or large in relation to the observed mean or other statistic, or a percentage of error that is small or large, does not suffice to tell us whether we are in the presence of a Salmonian coincidence or not, *without some sort of specification of the a priori range of numerical possibilities based on our background knowledge.* This is strikingly seen in frontier fields of science such as cosmology, where astrophysicists are sometimes quite pleased when a prediction "fits" within an order of magnitude, a 1,000% error being accepted as corroborative! This seems absurd until one takes account of the fact that the a priori range of cosmological big numbers is vast. Likewise, it would be corroborative of molecular theory if it predicted a value for Avogadro's constant at $6 \times 10^{23}$ and an experimental result gave us, say, $3 \times 10^{22}$. If we got a half dozen experimental values distributed anywhere around order of magnitude 23, we would consider first that some of the auxiliaries must be poor approximations (although not qualitatively false). If that Lakatosian retreat did not work, we would consider the theory falsified *as it stands.* Having given us a half dozen very strange coincidences as to order of magnitude, we would appraise it as worth retaining for amendment. The point is that there is no way to assess a standard error expressed in original units, or as a pure number canceling out the physical units, without some background knowledge giving us an idea, however rough, of the a priori range of possible values. I think the history of the developed sciences shows that this kind of thing happens over and over again and is such a matter of course that it is not even discussed as an epistemological point, being simply covered under the heading of such everyday scientist language as "reasonably accurate prediction." The notion of accuracy, when pressed, is a *relative* term, usually uninterpretable with respect to theory corroboration without the a priori range. The problem is that the concept of the a priori range and the concept of background knowledge are fuzzy concepts and therefore unsatisfactory if we are epistemological perfectionists. All I can say is that here again, as in the case of the verisimilitude concept, we have to do the best we can, because we simply can't do without it.

If I tell you that a measurement has a standard error of so many angstroms, you don't know how accurate that is without knowing something of the range of values we are concerned with in the particular experimental domain. If I tell you that a certain measurement was 1,000 miles off, you will think poorly of it if we are talking about terrestrial geography; you will be somewhat critical if we are talking about the average distance to the moon (an error of 0.4%); and you will consider it a minuscule error when dealing with the distance of our sun from Alpha Centauri. If I tell you that I have a genetic theory that enables me, from studying the biochemistry of the parents, to predict the length of a baby elephant's trunk with an average error of an inch, what do you make of this? You don't know what to make of it in appraising my genetic theory unless you know something about the range of trunk lengths in neonatal elephants. I won't belabor the point with other examples, because it's blindingly obvious, despite the fact that sometimes we have difficulty in saying what range the background knowledge plausibly allows.

It is sometimes possible in fields employing statistics to specify the theoretically possible range on mathematical grounds, if we are given a portion of the empirical data and asked to predict the rest of it. I take a simple example, a degenerate case of path analysis in testing a causal theory. Imagine a city with endemic cholera in which sewage is discharged into a canal that runs through the city, and the water supply comes from the canal. Some households and some hotels, for reasons of taste, snobbery, or suspicions about health, do not drink the canal water supply, but purchase bottled water. Some living on the outskirts of the city, where there are plentiful springs, get their drinking water from the springs. Because of location and expense, there is a statistical relationship between income and canal water consumption, but there are many exceptions. For example, the families living at the outskirts, near the springs, tend to be lower-middle class; center-city people are mostly lower-middle and lower class; but there are some fancy hotels in the middle of the city which regularly use the city water supply, but do make bottled water available for those guests who are willing to pay extra for it. It is known from clinical experience of physicians and common observation that poor people have more cholera, and it is also well known that poor people drink more canal water. One epidemiologist has a theory that cholera is due to a specific etiological agent found in the canal water and not otherwise transmitted, and he believes that poverty as such has no direct causal influence on cholera incidence. Another epidemiologist thinks that, although there may be something to the canal water theory, poverty predisposes to cholera by a combination of causal influences such as poor diet, crowded living conditions, poor hygienic practices, and psychosomatic stress lowering one's resistance to disease. Suppose these two epidemiologists know only the correlation coefficients—the units of measurement being city blocks—between $x$ = the poverty index and $z$ = canal water consumption ($r_{xz}$ = .60) and between $z$ = canal water consumption and $y$ = cholera incidence ($r_{zy}$ = .90) They each try to predict the correlation coefficient between poverty and cholera ($r_{xy}$). From the conventional path analyst's point of view this is an unsatisfactory epistemic situation because the path diagram is just barely determined, so we would be likely to say "no good test." But a Popperian would be less pessimistic, recognizing that the conventional path analyst is requiring a *deduction* when insisting that the system must be overdetermined, and we do not ordinarily require a deduction from facts to theory in empirical science, for the very good reason that none such can exist! The Popperian point here is that the first epidemiologist who believes in the specific etiology of cholera and accordingly thinks that the only reason poverty and cholera are related is that poverty has a causal path running through canal water consumption, would predict that the partial correlation $r_{xy.z} = 0$, which leads directly from partial correlation algebra to the prediction that $r_{xy}$ = .54, a point prediction that the other epidemiologist cannot make because his causal theory does not give rise to an empirical prediction one way or another. Neither theory is refuted by these results, but the second theory has to be tailored ad hoc to fit the results, which it could not have predicted in advance; whereas the first theory, that the only relationship between poverty and cholera incidence is causally mediated by canal water consumption, generates a point prediction, which turns out to be empirically correct.

What is the a priori range of possibilities here? One could argue that because we are talking about correlation coefficients, the possibilities range from –1 to +1, but that is not true when we are given the first two correlations as presented to both of our theorists. The partial correlation formula leads to a theoretically possible range for $r_{xy}$ which we get by writing the inequality $-1 \leq r_{xy.z} \leq +1$, an algebraic truth about the Pearson $r$ that is free of the usual assumptions such as normality and homoscedasticity or, for that matter, even rectilinearity. (The formula for partial correlation, although based on correlating the residuals around straight lines, does not require that the straight line be the best fit, i.e., that the correlation coefficient should be the appropriate descriptive statistic; rather, these formulas go through as a matter of sheer algebra.) Solving on both sides of the inequality we find that given the first two correlation coefficients, the a priori range of numerically possible values for the to-be-predicted $r_{xy}$ is between +.19 and +.90. Applying the principle of indifference, as the first epidemiologist's prediction is on the nose at $r_{xy}$ = .54, we have picked out 1 of 71 intervals on a rectangular distribution, a strange coincidence to the extent of $p < .02$. Although this reasoning looks like the traditional flabby significance test, it is of course much stronger than that, because it asks how likely it would be by chance not merely that there would be more cholera among the poor, but that the correlation between poverty index and cholera would be picked out of the a priori range with this accuracy.

This focusing on the size of the predicted interval in relation to an a priori range of numerical possibilities bears on an article by Hedges (1987). His important contribution helps to soften the Popperian blow to social scientists and should relieve some of their inferiority complexes with respect to fields like astronomy, physics, and chemistry. But one must be careful not to let it blunt the Popperian critique and lull us into unwarranted satisfaction. Hedges's treatment, epistemologically and mathematically sophisticated as it is, I do not criticize here. But he did not find it necessary for his clarification to make explicit how numerical tolerances in the developed sciences relate to the a priori range of possibilities, the point I am here emphasizing. One may, for instance, have good reasons, either from theoretical knowledge of experimental weaknesses or from a study of the obtained distribution of values, for excluding what to a conservative Fisherian psychologist would be an excessively large fraction of numerical outliers. Nevertheless, it could still be true (and would typically be true in fields like physics) that the change thereby induced in a statistical estimator of some physical constant would be small in relation to the a priori *conceivable* range of values that one might contemplate as possible, without the substantive theory. Furthermore, as Hedges himself pointed out, there is a difference between experiments aimed at determining a physical constant as accurately as possible, where it may be rational to exclude outliers, and experiments in which a numerical value is being employed to *test* the substantive theory. In the one case we have already corroborated the theory in a variety of ways, and we have quite accurate knowledge of the other physical constants relevant to our particular experiment. Our aim in excluding outliers is to reduce the standard deviation of the measures and hence the standard error in estimating the parameter (and probably a bias in the mean due to "gross error" in the excluded outliers), the theory in which all this numerical reasoning is embedded being taken as unproblematic. That is different from the typical situation in psychology where our estimate of a numerical value, or our refutation of the null hypothesis, is being taken as evidence for or against the substantive theory, which is in doubt. Testing a theory via a predicted numerical value, or (weakly but still quite satisfactorily) by the coherence of numerical values within small tolerances, is epistemically a different situation from the kinds of examples Hedges addresses in his article.

Let the expression *Lakatosian defense* designate the strategy outlined by Lakatos in his constructive amendment of Popper, a strategy in which one distinguishes between the hard core of $T$ and the protective belt. In my notation Lakatos's protective belt includes the peripheral portions of $T$, plus the theoretical auxiliaries $A_t$, the instrumental auxiliaries $A_i$, the ceteris paribus clause $C_p$, the experimental conditions $C_n$, and finally the observations $O_1$, $O_2$. The Lakatos defense strategy includes the negative heuristic which avoids (he said *forbids)* directing the arrow of *the modus tollens* at the hard core. To avoid that without logical contradiction, one directs the arrow at the protective belt.

However, Lakatos treated the defense as aiming to preserve *the literal truth of the hard core of T,* whereas I am softening that to say that we are merely adopting the weaker position that *the hard core of T has high verisimilitude.*

The tactics within the Lakatosian defensive strategy may vary with circumstances. As mentioned earlier, we may refuse to admit the falsifying protocol into the corpus, or raise doubts about the instrumental auxiliary, or challenge the ceteris paribus clause, or the theoretical auxiliaries, or finally, as a last ditch maneuver, question the peripheral portions of the substantive theory itself. Nobody has given clear-cut *rules* for which of these tactics is more rational, and I shall not attempt such a thing. At best, we could hope to formulate rough guidelines, rules of thumb, "friendly advice," broad *principles* rather than *rules* (Dworkin, 1967). It is easy, however, to make some plausible suggestions. For instance, if the fact domain is readily divisible into several qualitatively different experimental contexts, and one finds a piling up of falsifiers in one of them, it would seem reasonable to challenge the ceteris paribus clause there, rather than amending auxiliaries, which cut across the subdomains. If the theory is quantitative, altering an auxiliary to take care of a falsifier in one domain will, if that auxiliary appears in other domains as well, generate falsifications in them, because the data that fitted the original auxiliary mathematical function will now, curve-fitting problems aside, no longer fit them. With regard to the decision whether to admit the falsifying protocol into the corpus, that can depend on the previous track record of the experimenter as to replicability of findings reported from a particular laboratory, the adequacy with which the experimental setup was described, and the like. These are fascinating and important questions in which little progress has been made so far by the philosophers of science, and 1 shall say no more about them here. The main point is that conducting a Lakatosian strategic defense, whichever aspects of the protective belt we focus on in our positive heuristic, is not predicated on belief that in the long run the hard core of *T* will turn out to be literally true (although that may be included as one of the optimistic possibilities), but rather on our conjecture that the hard core of *T* will turn out in the long run to have possessed high verisimilitude. Of course, to the extent that we apply the positive heuristic to the auxiliaries and ceteris paribus clause, rather than making inroads into the peripheral portions of *T* itself, we are reasoning temporarily *as if* the literal truth of *T*, both hard core and periphery, might obtain.

When is it rational strategy to conduct a Lakatosian defense? Here we invoke the Lakatos principle. We lay down that it is not a rational policy to go to this much trouble with amendments of *T* or adjustments of auxiliaries unless the theory already has money in the bank, an impressive track record, and is not showing clear symptoms of a degenerating research program.

How does a theory get money in the bank—how does it earn an impressive track record? We rely on the basic epistemological principle that "If your aim is a causal understanding of the world, do not attribute orderliness to a damn strange coincidence." We could label this "Reichenbach's maxim," because in his famous justification of the straight rule of induction he says that, although we can have no guarantee it will work, it will work if anything works. Or we might label it "Novalis's maxim," remembering the epigraph of Popper's great 1935 book, quoted from Novalis, "Theories are nets: Only he who casts will catch." We apply this maxim to formulate Salmon's principle: that the way a theory gets money in the bank is by predicting observations that, absent the theory, would constitute damn strange coincidences. I don't label this "Popper's principle," because accepting the Serlin–Lapsley critique of my overly Popperian earlier statements, I am here emphasizing that a theory can get a lot of money in the bank, and hence warrant us in conducting a Lakatosian defense, despite its being falsified. It does this by achieving a mixture of *risky successes (*passing strong Popperian tests) and *near-misses,* either of these being Salmonian damn strange coincidences.

## $H_0$ Testing in Light of the Lakatos–Salmon Principle

How does the conventional null-hypothesis refutation procedure fare under the aegis of the joint Lakatos–Salmon principle? As a start, let us set aside the purely statistical problem, which receives almost all the emphasis in statistics classes, by assuming that we have perfectly valid measures and no sampling error because (a) there are no appreciable individual differences, or (b) we have exhausted the physically specified population, or (c) we have such a gigantic *N* that sampling error is negligible. Now suppose we have performed 10 experiments (or 10 statistical studies of our clinical file data) predicting in each case from our weak theory that one mean will be higher than the other. Assume that the 10 experiments are in highly diverse qualitative domains, as with the Perrin determinations of Avogadro's number, so that they can be treated as experimentally and statistically independent, although of course they are not conceptually so in the light of the theory being tested. Having heard of Popper, and being aware that the formal invalidity of the third figure of the implicative syllogism is dangerous in the empirical realm, we set up a fairly strict significance level of alpha = .01. To reach that level in 10 experiments, 9 must come out in the expected direction. If we have a couple of dozen experiments, around three fourths of them have to come out in the expected direction; if we have as many as 50 independent experiments, between two thirds and three fourths must do so. Anyone familiar with narrative summaries of research in the soft fields of psychology (and often even in the "hard" ones) knows that these box-score requirements are not likely to be met.

Now contrast this situation with 10 narrow-range or point predictions as in the Avogadro problem. Performing even two experiments making such precise predictions yields $p$ = .01 if the subintervals within the a priori range are as small as one tenth, because the probabilities are multiplied. Because these probability products go up exponentially, null-hypothesis testing is much feebler because what *it* tells us is merely that a given testing will fall in the upper rather than the lower half of the a priori numerical range.

This obvious comparison answers one defense of the conventional method that I hear from students and colleagues who are made nervous by the Popperian critique of feeble theory testing by significance tests, in which they point out that a significance test can be restated in the form of an interval estimation despite Fisher's (1925, 1937) strong emphasis on the difference between the two problems. The mathematics is identical, and instead of saying that I have refuted the point $H_0$ at level alpha (especially considering that point $H_0$ is always false in the life sciences, so whether we succeed in refuting it simply depends on the statistical power function) I could use the same algebra to make the statement that I have a probability of .95 that the difference lies on the positive side of zero. The confidence-interval equivalent of a directional $H_0$ refutation is large, typically around one half, so that the joint (multiplicative) probability of several "successful outcomes" does not fall off nearly as rapidly as happens when one makes a numerical prediction of a point value or a small interval.

For instance, let us say we have a causal theory about the influence of genes and home environment, and the relative importance of father and mother as caregivers and intellectual stimulators; but the theory is so weak that it merely predicts that a foster child's IQ will be somewhat closer to that of the foster mother than to the IQ of the foster father. A finding in that direction (again assuming away sampling error and imperfect measurement) has an even chance of being right, whether or not our theory has any verisimilitude. Whereas if we have a strong enough genetic model to make point predictions of IQ values, hitting the correct value within a point or two already has a fairly low prior probability absent the theoretical prediction.

But matters are worse than this, for a nonstatistical reason. Even if a batch of null-hypothesis refutations is piled up enough in one direction to generate a small conjoint chance probability, that provides only rather feeble corroboration to a *substantive* theory *T*. When we avoid the seductive tendency to conflate *T* with a directional statistical hypothesis $H^*$ (by which I mean the opposite of the directional null hypothesis of zero or negative difference), what does a small probability of a pileup of directional findings corroborate? All it corroborates is the "theory" that *something nonchance must be at work in one direction.* As Dar (1987) pointed out in his reply to Serlin and Lapsley (1985), that is not a very strong finding. There is a pretty big class of actual and possible *T*s easily capable of generating a directional expectation along these lines. Thinking Bayesian, that amounts to pointing out that, in the denominator of Bayes's theorem, the expectedness has two components, the second of which is the sum of the products of the prior probabilities on all the competitor theories capable of generating this same kind of directional fact by the conditional probabilities of a directional finding.

More sophisticated readers may suppose that I am here beating a dead horse, that every thoughtful social scientist surely knows about the reasoning in the preceding paragraphs, but that is simply not true. As an example, I recently heard a colloquium in which the investigator was interested in the effect of childhood sexual abuse on the sexual and self-concept attitudes of college males. A set of about a dozen adult attitude and experience characteristics were the presumed causal "output." Only three or four of these output measures were statistically significant, and because the statistical power of his *N* was pretty good, one must view the batting average as poor. (Note that if the theory predicts effects on all these output measures—he would doubtless have counted them as "support" had they panned out!—we must describe it as refuted.) Of course he focused his attention on the ones that did show a difference, but made no mention of the effect sizes. When I asked in the discussion period roughly how big were the effects, he said he didn't know! In fact, his table showed them to be around a half standard deviation, which would mean that if one located the *hitmax* cut (Meehl, 1973a) midway between the abused and nonabused means on the (selected) subset of outcome measures that reach statistical significance, and tried to predict a pathological adult attitude or practice on the grounds of knowing the subject had been sexually abused as a boy, the normal curve tables indicate that one would do around 10% better than by flipping pennies.

All sorts of readily available theories based not on ad hockery but on the research literature are easy explainers of such a small trend as this. There might be differences in repression of childhood events; differences in self-revelation willingness; the MMPI *K* factor present in all inventories; possible factors of introspection, intelligence, verbal fluency, social class, and the like. Any one (or more) of these could be correlates of genetic loadings for the subset who were abused by biological relatives, which same genetic loadings might affect the sexual behavior and self-concept of the abused subjects as college adults, and so on and on ....

The point is that finding a difference of this size is a feeble corroborator of the etiological relation that the research was supposed to be about. It testifies to the stupefaction induced by conventional statistics training that this researcher, having run his *t* tests, was not even curious enough to look at the effect sizes! I would have been embarrassed had a professor of physics, chemistry, or genetics been in that audience.

## The Crud Factor

Research in the behavioral sciences can be experimental, correlational, or field study (including clinical); only the first two are addressed here. For reasons to be explained (Meehl, 1990c), I treat as correlational those experimental studies in which the chief theoretical test provided involves an interaction effect between an experimental manipulation and an individual-differences variable (whether trait, status, or demographic). In correlational research there arises a special problem for the social scientist from the empirical fact that "everything is correlated with everything, more or less." My colleague David Lykken presses the point further to include most, if not all, purely experimental research designs, saying that, speaking causally, "Everything *influences* everything," a stronger thesis that I neither assert nor deny but that I do not rely on here. The obvious fact that everything is more or less correlated with everything in the social sciences is readily foreseen from the armchair on common-sense considerations. These are strengthened by more advanced theoretical arguments involving such concepts as genetic linkage, auto-catalytic effects between cognitive and affective processes, traits reflecting influences such as child-rearing practices correlated with intelligence, ethnicity, social class, religion, and so forth. If one asks, to take a trivial and theoretically uninteresting example, whether we might expect to find social class differences in a color-naming test, there immediately spring to mind numerous influences, ranging from (a) verbal intelligence leading to better verbal discriminations and retention of color names to (b) class differences in maternal teaching behavior (which one can readily observe by watching mothers explain things to their children at a zoo) to (c) more subtle—but still nonzero—influences, such as upper-class children being more likely Anglicans than Baptists, hence exposed to the changes in liturgical colors during the church year! Examples of such multiple possible influences are so easy to generate, I shall resist the temptation to go on. If somebody asks a psychologist or sociologist whether she might expect a nonzero correlation between dental caries and IQ, the best guess would be yes, small but statistically significant. A small negative correlation was in fact found during the 1920s, misleading some hygienists to hold that IQ was lowered by toxins from decayed teeth. (The received explanation today is that dental caries and IQ are both correlates of social class.) More than 75 years ago, Edward Lee Thorndike enunciated the famous dictum, "All good things tend to go together, as do all bad ones." Almost all human performance (work competence) dispositions, if carefully studied, are saturated to some extent with the general intelligence factor *g,* which for psychodynamic and ideological reasons has been somewhat neglected in recent years but is due for a comeback (Betz, 1986).

The ubiquity of nonzero correlations gives rise to what is methodologically disturbing to the theory tester and what I call, following Lykken, the *crud factor.* I have discussed this at length elsewhere (Meehl, 1990c), so I only summarize and provide a couple of examples here. The main point is that, when the sample size is sufficiently large to produce accurate estimates of the population values, almost any pair of variables in psychology will be correlated to some extent. Thus, for instance, less than 10% of the items in the MMPI item pool were put into the pool with masculinity–femininity in mind, and the empirically derived *Mf* scale contains only some of those plus others put into the item pool for other reasons, or without any theoretical considerations. When one samples thousands of individuals, it turns out that only 43 of the 550 items (8%) fail to show a significant difference between males and females. In an unpublished study (but see Meehl, 1990c) of the hobbies, interests, vocational plans, school course preferences, social life, and home factors of Minnesota college freshmen, when Lykken and I ran chi squares on all possible pairwise combinations of variables, 92% were significant, and 78% were significant at $p < 10^{-6}$. Looked at another way, the median number of significant relationships between a given variable and all the others was 41 of a possible 44. One finds such oddities as a relationship between which kind of shop courses boys preferred in high school and which of several Lutheran synods they belonged to!

The ubiquity of the crud factor is what gave rise to the bizarre model I propounded in my 1967 article against null-hypothesis testing, in which an investigator draws pairs of variables randomly from an empirical variable hat, and draws theories randomly out of a theory hat, associating each theory with a pseudopredicted empirical correlation. Due to the crud factor, that investigator would come up with a sizable number of apparent "substantiations" of the theories even if they had negligible verisimilitude and there were no intrinsic logical connections between the theory and the pair of variables employed for "testing" purposes.

I find three objections to this model from defenders of the conventional null-hypothesis approach. One objection is that no investigator would proceed in such a crazy way. That misses the point, because this irrational procedure is the worst scenario for getting a favorable ("theory-supporting") result, and my argument is that even in this absurd situation one can expect to get an encouraging number of pseudocorroborations of the theory. Just how many will depend jointly on (a) the average size of the crud factor in a particular research domain and (b) the value of the statistical power function.

A second objection is against treating such a vaguely defined class of actual and possible theories as a statistical collective, and the associated reliance on the principle of indifference with respect to directionality. To this objection I reply that if one is unwilling to consider a vaguely defined class of actual and possible experimental setups, then one would be unable to apply the probability values yielded by a significance test for interpretive purposes, that is, to apply Fisherian thinking itself. If a significance test is to permit an inference regarding the probative value of an experiment, it always implicitly refers to such a hypothetical class. One of the clearest examples where the principle of indifference is acceptable to logicians and statisticians is the case in which the procedure itself is a randomizing one, which is Fisher's preferred definition of the concept of randomness (i.e., 'randomness' referring not to the *result,* but to the *procedure;* this distinction lies behind Fisher's objection to the Knut Vik square in agronomy).

The third objection is somewhat harder to answer because it would require an encyclopedic survey of research literature over many domains. It is argued that, although the crud factor is admittedly ubiquitous—that is, almost no correlations of the social sciences are literally zero (as required by the usual significance test)—the crud factor is in most research domains not large enough to be worth worrying about. Without making a claim to know just how big it is, I think this objection is pretty clearly unsound. Doubtless the average correlation of any randomly picked pair of variables in social science depends on the domain, and also on the instruments employed (e.g., it is well known that personality inventories often have as much methods-covariance as they do criterion validities). A representative pairwise correlation among MMPI scales, despite the marked differences (sometimes amounting to phenomenological "oppositeness") of the nosological rubrics on which they were derived, is in the middle to high .30s, in both normal and abnormal populations. The same is true for the occupational keys of the Strong Vocational Interest Blank. Deliberately aiming to diversify the qualitative features of cognitive tasks (and thus "purify" the measures) in his classic studies of primary mental abilities ("pure factors," orthogonal), Thurstone (1938; Thurstone & Thurstone, 1941) still found an average intertest correlation of .28 (range = .01 to .56!) in the cross-validation sample. In the set of 20 California Psychological Inventory scales built to cover broadly the domain of (normal range) "folk-concept" traits, Gough (1987) found an average pairwise correlation of .44 among both males and females. Guilford's Social Introversion, Thinking Introversion, Depression, Cycloid Tendencies, and Rhathymia or Freedom From Care scales, constructed on the basis of (orthogonal) factors, showed pairwise correlations ranging from –.02 to .85, with 5 of the 10 $r$s $\geq .33$ despite the purification effort (Evans & McConnell, 1941). Any treatise on factor analysis exemplifying procedures with empirical data suffices to make the point convincingly. For example, in Harman (1960), eight "emotional" variables correlate .10 to .87, median $r = .44$ (p. 176), and eight "political" variables correlate .03 to .88, median (absolute value) $r = .62$ (p. 178). For highly diverse acquiescence-corrected measures (personality traits, interests, hobbies, psychopathology, social attitudes, and religious, political, and moral opinions), estimating individuals' (orthogonal!) factor scores, one can hold mean $r$s down to an average of . 12, means from .04 to .20, still some individual $r$s > .30 (Lykken, personal communication, 1990; cf. McClosky & Meehl, in preparation). Public opinion polls and attitude surveys routinely disaggregate data with respect to several demographic variables (e.g., age, education, section of country, sex, ethnicity, religion, education, income, rural/urban, self-described political affiliation) because these factors are always correlated with attitudes or electoral choices, sometimes strongly so. One must also keep in mind that socioeconomic status, although intrinsically interesting (especially to sociologists) is probably often functioning as a proxy for other unmeasured personality or status characteristics that are not part of the definition of social class but are, for a variety of complicated reasons, correlated with it. The proxy role is important because it prevents adequate "controlling for" unknown (or unmeasured) crud-factor influences by statistical procedures (matching, partial correlation, analysis of covariance, path analysis).

The crud factor is only 1 of 10 obfuscating factors that operate jointly to render most narrative summaries of research in soft psychology well-nigh uninterpretable. These 10 factors are:

1. Loose (nondeductive) derivation chain, making several "obvious" inferential steps requiring unstated premises (intuitive, common-sensical, or clinical experience).
2. Problematic auxiliary theories, although explicitly stated.
3. Problematic ceteris paribus clause.
4. Imperfect realization of particulars (experimenter mistakes in manipulation) or experimenter bias in making or recording observations.
5. Inadequate statistical power to detect real differences at the conventional significance level.
6. Crud factor: In social science everything correlates with everything to some extent, due to complex and obscure causal influences.
7. Pilot studies used to (a) decide whether "an effect exists" and (b) choose a sample size of adequate statistical power if the pilot effect is borderline but in the "right direction."
8. Selective bias in favor of submitting reports refuting the null hypothesis.
9. Selective bias by referees and editors in accepting papers refuting the null hypothesis.
10. Detached validation claim for psychometric instruments.

Factors 1 to 5 tend to make good theories look bad. Factors 6 to 9 tend to make bad theories look good. Factor 10 can work either way. Because these 10 obfuscators are usually nonnegligible, of variable and unknown size, and mutually countervailing, rational interpretation of an empirical "box score" is difficult—I would say typically impossible. Detailed treatment of these obfuscators and their joint quantitative influence is found in Meehl (1990c). Focusing on the obfuscator that is least recognized by social scientists, I provide one simple numerical example to illustrate the point that a modest crud factor cannot be discounted in the metatheory of significance testing. Returning to our absurd model of the fact hat and the theory hat, suppose that a representative value of the crud factor in a certain research domain were $r$ = .30, not an implausible value from the examples given. We have a substantive theory $T$, and we are going to "test" that theory by a correlational study involving observable variables $x$ and $y$, which, however, have no intrinsic logical connection with $T$ and have been drawn randomly from our huge pot of observables. Assume both $x$ and $y$ are approximately normal in distribution. We dichotomize the independent variable $x$ at its mean, classify each subject as high or low on the $x$ trait, and compare their scores on the dependent variable $y$ by a $t$ test. With the mean standard score of the highs on $x$ being .8 (at +1 *MD)* and that of the lows being –.8, there is a difference of 1.6 sigma in their means. Hence the expected mean difference on the output variable is $d$ = .48, about half a sigma. Assuming sample sizes for the highs and lows are around 37 (typical of research in the soft areas of psychology), we find that the probability of reaching the 5% level in a directional test is .66. So a theory that has negligible verisimilitude, and where there is no logical connection between the theory and the facts, has approximately a 2-to-1 chance of being corroborated provided that we were predicting the correct direction. If one assumes that the direction is completely chance (which in any real research context it would not be, for a variety of reasons), we still have a .33 probability of squeaking through with a significant result; that is, the empirical probability of getting a positive result for the theory is larger, by a factor of 6 or 7, than the .05 we have in our minds when we do a $t$ test. There is, of course, nothing wrong with Fisher's mathematics, or the tables. It's just that they tell us what the probability is of obtaining a given correlation if the true value is zero, whereas what we need to know, in appraising our theory, is how the correlation stands in relationship to the crud factor if the theory were false.

*The crud factor is not a Type I error.* It is not a statistical error at all. The crud factor refers to real (replicable) correlations which, although themselves subject to sampling error, reflect true causal relationships among the entities under study. The problem is methodological, not statistical: There are too many available and plausible explanations of an $xy$ correlation, and, besides, these explanations are not all disjoint but can often collaborate. Some minitheories are objectively of high verisimilitude, including theories that nobody gets around to formulating. The observed distribution of correlation coefficients among all the observable variables in a certain domain, such as the hundreds of different personality traits for which various measures exist, are a consequence of certain real causal factors. They have their explanation in the grand theory $T_{OJ}$ known to Omnisicient Jones but not to us. The problem with null-hypothesis refutation is that to the extent that it corroborates anything, it corroborates the whole class of theories capable of generating a nonzero directional difference. There are simply too many of them in soft psychology for this to constitute a distinctive test. The bite of the logician's point about "affirming the consequent" being in the third figure of the implicative syllogism lies in the number of different ways that the consequent might be entailed. In soft psychology this number is unknown, but it is certainly not small.

To make this less abstract, I give some psychological examples. Suppose we test my theory of schizotaxia (Meehl, 1962, 1989, 1990b, 1990d) by running the Whipple steadiness test on the first-degree relatives of schizophrenes. Briefly, the theory postulates a dominant schizogene which produces a special sort of synaptic slippage throughout the central nervous system (CNS), giving rise in the endophenotype to a neural integrative defect, giving rise in the exophenotype to multiple soft neurology and psychophysiology indicators. Suppose we find that the first-degree relatives of schizophrenes manifest a deficient motor steadiness. How strongly does this corroborate my theory? Weakly, although not zero. Several alternative explanations spring to mind readily, and I doubt it would take a graduate student in psychology more than five minutes to come up with a half dozen or more of them. Alternative plausible hypotheses include:

1. The subjects know, or easily infer, that they are the subjects of study because they have a schizophrenic relative and are made anxious (and hence tremulous) by wondering what the experimenters are thinking of them.
2. The subjects are not worried about the experimenter's opinion but have at times had doubts as to their own mental health and worries as to whether they might develop schizophrenia, and this experimental setting mobilizes those anxieties.
3. Contrary to Meehl's theory, schizophrenia is not genetic but is due to the bad child-rearing practices of a schizophrenogenic mother; although she damages the proband more than the siblings, they were also exposed to this environment and consequently they have a generalized tendency to heightened anxiety and, hence, motor tremor.
4. Schizophrenia is heritable but not neurological. Rather, polygenic variables affect the size of the anxiety parameter, and the subjects were fortunate enough to get somewhat fewer anxietous polygenes than the proband, but enough to make them different from the controls.
5. The theory is correct in conjecturing something subtle about CNS function, and the soft neurology in psychophysiology are consequences of this rather than emotional factors as in the previous examples, but they do not involve a major locus.
6. Soft neurology and social anxiety are pleiotropic indicators of the schizogene, the latter not being mediated at all in the way Meehl conjectures.

Suppose one has half a dozen such plausible conjectures to account for the existence of a nonzero difference between the relatives and controls. Without any basis for preferring one to the other, if you plug the positive experimental result into Bayes's formula you find that each theory's posterior probability given the successful outcome is .16, even assuming that your list of possibilities is exhaustive—which it is not. A strong test will involve taxometric methods (Meehl & Golden, 1982) of proving, first, that a subset of the first-degree relatives represents a taxon; second, that the base rate of that taxon among parents and siblings is close to the $P = 1/2$ required by the dominant-gene conjecture; and, finally, that one member of *each* parent pair must belong to the taxon, from which follows some further quantitative statistics about their scores (Golden & Meehl, 1978). For another example involving schizophrenia theory, see my discussion of alternative causal chains resulting in lower high-school social participation by preschizophrenes (Meehl, 1971).

Or consider the famous "pratfall" experiment of my friend and former colleague Elliot Aronson and his co-workers (Aronson, Willerman, & Floyd, 1966). I choose this one because it is a cute experiment and because the theoretical conjecture is an interesting one, unlike many of those in personality and social psychology which are trivial, being common-sense truths (Leon Festinger called it "bubba" psychology, for "what my grandmother knew") formulated in pedantic language. I don't wish to dispute Aronson's theoretical interpretation but only to suggest how easy it is to cook up possibilities. The finding was that when one has positive prestigeful evaluations of a person who commits a social gaffe or blooper in a public setting, this results in a shift in favorable attitude toward the victim. (I set aside the size of the difference, which in the soft fields of psychology is almost never considered, or even reported. This business of "Jones showed that $x$ is related to $y$" or, more offensive to one who knows anything about the powerful sciences, "Smith showed that $x$ is a function of $y$" is a bad habit in reporting social science research.) What are some of the theoretical possibilities?

1. Thinking psychodynamically, we might suppose that, if the victim is a prestigious figure in my value system, I will feel unconscious hostility because of my competitive impulses, which I will have to defend against, say, by reaction formation, which will lead me to make positive ratings.
2. I identify with this prestige figure, and, because I would wish to be treated nurturantly in case of such a slip, I treat the victim nurturantly in my postslip evaluation.
3. I do not identify with or feel competitive toward him, but the whole situation strikes me as amusing, and, when I feel amused, I tend to feel broadly "positive" about anybody or anything.
4. The initial situation threatens me competitively, but his slip "brings him down to my level," so I feel relieved, and increments in hedonic tone tend diffusely to influence momentary plus/minus evaluations.
5. I feel guilty at my flush of pleasure over his discomfiture, and the defense mechanism activated is undoing rather than reaction formation.
6. Finally, we have the conjecture propounded by Aronson and his co-authors: that the blunder "humanizes" him, increasing his attractiveness. (Is this identical with my fourth possibility, or distinguishable?)

An abstract way to get an appreciation of this problem is to reflect on the number of theoretical variables available for explaining observed correlations in the soft areas. If the psychisms mobilized result from personality traits (activations of dispositions), screenings beginning with the 18,000 trait names in the famous Allport-Odbert (1936) list have rarely succeeded in reducing the number of distinguishable and in some sense "important" traits to less than 100 (see, e.g., Meehl, Lykken, Schofield, & Tellegen, 1971; Meehl et al., 1962). Of course these are surface traits, and one might prefer to invoke source traits ("genotypic traits," dispositions to internal and not always conscious psychisms) before counting it as a real explanation. A simple configuration is the triad provided by a Murray need, a mechanism of defense ("defense" here used loosely to mean any method of handling the need, whether or not in the interest of avoiding the anxiety signal in Freud's sense), and one of a set of objects. In research I was engaged in many years ago, we narrowed the list of Murray needs down to around 20, the list of defense mechanisms to around the same number, and provided the therapists making ratings with a set of some 30 objects

(Meehl, 1964). Theoretically this would give us 400 need–defense combinations. If we say that only a minority of possible objects are candidates for a given need (say, as few as 10%), we still have more than 1,000 need–defense–object triadic patterns to deal with. If, to explain a particular correlation or experiment, I can without Procrustean forcing plug in either of 2 needs, 2 defenses per need, and then choose among 3 objects, I still have 12 possible minitheories, giving a posterior probability of only .08 assuming equal Bayesian priors. The methodological situation here is well expressed by cynic Ring Lardner's maxim, "In general, the odds are 8 to 5 against." Researchers in the soft areas who are sensitized to this inferential problem would presumably expect to perform a *minimum* of 12 experiments to exclude competing minitheories, a practice which, so far as I am aware, no investigator follows.

One might say, "Well, what about chemists? They have all these chemical elements to worry about." Yes, and they have specific tests that exclude whole classes of them in performing a qualitative analysis; and they supplement qualitative analysis with quantitative analysis when necessary to rule out other possibilities; and there are alternative high-validity indicators (e.g., chemical reagents, chromatography, spectroscopy) that cohere in their indications, as in the Avogadro case. Even in the study of animal learning and motivation, a simple dispositional analysis operating with a model like Carnap's (1936–1937) reduction sentences becomes complicated in a hurry, because testing one disposition by a certain reduction sentence will involve ceteris paribus clauses about other variables which in turn have to be subjected to exclusion tests, and so on. (Cf. Skinner, 1938, p. 25, on deciding whether the rat is extinguished, satiated, or afraid—a paradigm case of the psychologist's problem for a simple organism in a simple context.) The arch positivist Otto Neurath (1932–1933/1959) spoke of "repairing the raft you are floating on," and Popper (1935/1959) made the analogy to "sinking piles into a swamp." Unfortunately in the social sciences, the situation is more like standing on sand while you are shoveling sand (MacCorquodale & Meehl, 1954, pp. 232–234), and, alas, in soft psychology the sand is frequently quicksand.

Instead of the highly structured battery of experiments to rule out competitor minitheories, the typical researcher in soft psychology feels pleased by a box score that gives more successful than unsuccessful predictions, when these predictions consist of mere null-hypothesis refutations. The subset of predictions that come out "wrong"—which from a Popperian standpoint constitute strong falsifiers and, logically speaking, outweigh any preponderance of corroborators—are dealt with by ad hoc adjustments. These usually lead to doing another experiment on the ad hoc conjecture which, if it comes out positive, is considered a favorable result. If it doesn't, it is then adjusted, and so forth. This can give rise (as I pointed out in my 1967 article) to a sequence of experiments testing successive ad hoc adjustments, which, in the social climate of our field, gives one a reputation for carrying out a "sustained research program" but which, from Lakatos' standpoint, could often be taken to exemplify a degeneration.

A defender of the conventional approach might emphasize that the Popperian hurdle becomes higher, harder to surmount, a more powerful test, because the statistical power is imperfect. Agreed, but the price one pays for that is an increase of Type II errors, so the net effect of adding statistical inference problems to our imagined "error free" data pattern is to make the meaning of the box score even fuzzier than it already was. Because of the ineluctable trade-off between errors of Type I and Type II, the investigator is in danger of getting erroneous discorroborations of theories having high verisimilitude, and in soft psychology our problems of statistical power and methods-covariance make box scores well-nigh uninterpretable. Because the basic problem here is the weak *epistemic* linkage between $H$ and $T$, it is fruitless to try wriggling out of that difficulty by invoking the *statistical* slippage between $H$ and $O$. No statistical ingenuity can cure a logician's complaint about the third figure of the implicative syllogism, that the theory is a sufficient but not necessary condition for the fact, by casting doubt on the fact; that can only add insult to injury. As the sergeant major advised French Foreign Legion recruit John Smith, "When things are bad, *bleu,* do not make them worse, for they will be quite bad enough" (Wren, 1925).

### Appraising a Theory: Point and Interval Predictions

If one is persuaded by these considerations, the question arises whether one could roughly measure the Lakatosian status of a theory? Perhaps not, but I would like to have a try at it. I take a handy notion from the Vienna positivists (which they took, I believe, from Von Kries, a philosopher-statistician of the 19th century): the concept *of Spielraum* (German word for "action play," "play/game space," "field," "range," "scope," "elbow room"). In its original usage, relying on the principle of indifference this concept envisaged the range of logical possibilities. I am going to add to that way of arriving at it, a "background knowledge" way, as the Bayesians would say. In the earlier example of a simple path-analytic problem involving cholera and canal water, we fixed the Spielraum by combining two correlation coefficients with the algebra of partial correlation, plus the principle of indifference. Setting up a rough numerical Spielraum about a theory's predictions requires some sort of rational basis. Sometimes this is almost purely a priori; sometimes it involves considerable empirical background knowledge. However arrived at, the empirical context sets "reasonable" upper and lower bounds on a measured quantity, and we apply the principle of indifference, perhaps combined with purely formal considerations (as in the partial-correlation situation) to compute an a priori probability of being correct when we predict a point value or an interval. There is an unavoidable vagueness about this, *but it is in no worse shape than the epistemological vagueness provided by conventional significance testing.*

Here is one respect, however, in which the social sciences may have an advantage. By far the larger part of our research, when quantified, eventuates in relationships expressed by pure numbers, that is, where dimensional analysis of the quantification cancels out centimeters, dollars, IQ points, or whatever. Almost all the pure numbers we employ have algebraically defined bounds. The Pearson $r$ coefficient and its surrogates go from zero to one; analyses of variance and covariance are expressible in terms of proportion of variance accounted for; beta coefficients in a multiple-regression equation, the weights in a linear discriminant function, the factors in a factor analysis, the base rate and hit rates in taxometrics—all of which collectively comprise 90% of research in "soft" psychology—have mathematically defined ranges of possible values. In path analysis, we would have to adopt a convention as to whether the basic range of the reconstructed correlation should be employed as Spielraum, or, instead, the range allowed by the algebra of partial correlation given the data but not the path diagram.

In research areas involving physical units in which it is not customary to analyze the data in a way that eventuates in a dimensionless number, setting up suitable conventions would be harder and somewhat arbitrary. However, as long as we see clearly that the a priori range should not be based on the theory under test, reasonable rules of thumb could be arrived at. Thus, for example, if we are studying memory, the boundaries of the Spielraum could be taken simply as remembering everything and remembering nothing. If reaction time or the rate of responding in a cumulative record is the measure, and we are comparing two groups (or the same group before and after an intervention), it would be reasonable to say that the Spielraum goes from the highest value found in any individual in either group to the lowest value found in any individual in either group. So long as we do not entertain metaphysical absolutist ideas about what the index is attempting however crudely to quantify, a choice of convention for whole classes of experimental work need not be optimal as long as it's reasonable. As Mr. Justice Brandeis said, in many situations it is more important to have a rule than to have the best rule. If a construct-validity bootstrapsing based on factor analysis and discriminant analysis of several indices were carried out (as suggested in the discussion to follow) it is not a vicious circle to try out alternative Spielraum specifications in a given research domain, selecting the one that shows the highest factor loading when embedded in the multiple appraisal system.

To construct a crude index of a theory's track record, one first amends the earlier Popper to the later Popper by shifting emphasis from falsification to verisimilitude. Although at some stage of a research program the possibility of the core of $T$ being literally true may be seriously entertained, that would seem rare in psychology. But I suggest that this doesn't matter much strategically. Whether one looks on the Lakatosian defense as aimed (for the time being) at preserving a conjecture of perfect verisimilitude for the hard core, $T_{HC}$, or only defending the weaker conjecture that $T_{HC}$ has high verisimilitude, will not differentiate the early stages of a strategic Lakatosian retreat. We are assuming—despite the lamentable fact that no philosopher of science has provided a proof—that there is a stochastic relationship between a theory's track record and its verisimilitude (but cf. Meehl, 1990a). We wish to numerify that track record. I use 'numerify' as a more modest, neutral term than 'quantify,' which to some connotes *measurement,* and hence stronger claims about the metric than are possible or, for our purposes here, necessary. Numerifying is attaching numbers by rule, and may or may not claim strict ordination, interval or ratio scale, and so forth. Within such an approximative framework, the adages "a miss is as good as a mile" and "close, but no cigar" do not apply. A falsifying protocol, if admitted into the corpus, falsifies the conjunction on the left of our corroborative equation supra, leaving us considerable freedom in where to make the amendments. Meanwhile, we require of a candidate index that it somehow reflect *how bad* a numerical "miss" the experiment chalks up against *T. I am deliberately setting aside statistical significance testing, or the setting up of confidence intervals, whether used in the weak or the strong way.* We are examining the relationship between $T$ and its track record in predicting numerical values of $H$, ignoring the stochastic slippage between $H$ and the data set that is the main concern of the statistician.

Second, we require an index that does justice to the interesting fact that the working scientist is often more impressed when a theory predicts something within, *or close to,* a narrow interval than when it predicts something correctly within a wide one. Had I paid attention to this well-known fact, I would not have preached such a simplistic version of Popper in my earlier articles. Consider an example: On a conjectural causal model of the determiners of IQ, I predict the mean IQ of a defined group of children to be 117 ± 2. The data yield a mean of 120. For $\text{Popper}_0$ my theory is falsified. Does that mean I abandon it forthwith? Surely not. What do I say? "Well, it wasn't right on the nose, and strictly speaking it departed significantly from the allowed statistical tolerance around the predicted value, but by only one point. That's a fairly accurate value—a pretty close miss—*considering the range of possibilities a priori."* In contrast to this "close enough" situation, imagine a theory of intelligence so weak that it predicts merely that the IQ of a certain group ought to be above average. Cutting off at say, 3 sigma, the a priori Spielraum is from IQ 55 to IQ 145, so my weak theory has passed the test by correctly locating the observed mean in the upper half of this Spielraum. I cannot conceive that any psychologists would find this second *literally correct* result more exciting, giving the substantive theory more money in the bank, than they would the first one, where the prediction is off by 3 IQ points and the deviation exceeds the tolerance by one point. And there is nothing peculiar about psychology in this respect, it happens often in any science that uses quantitative methods. The crucial thing is, I urge, not the standard error, or even (somewhat more helpful) the engineer's familiar percentage error, but *the size of the error in relationship to the Spielraum.*

Even that doesn't give us all the information we want, as the IQ example shows. Closeness in relation to the Spielraum is one way to numerify Serlin and Lapsley's (1985) "good enough" principle. But given that, for a fixed size of error in relation to the Spielraum, we appraise a theory more favorably if its prediction was narrow with reference to the Spielraum. This is similar to Popper's original emphasis on corroboration being a function of risk, except that here again it is not yes-or-no falsification but Salmon's principle that we wish to numerify. The revised methodology retains the Popperian emphasis on riskiness, but now instead of asking "Did I pass the test, which was stiff?" we ask, "How close did I come?" The ideal case of strong corroboration is that in which the theory predicts a point value (a point value always means, in practice, an interval) and succeeds. A less favorable case, but still leading to a positive appraisal, is a theory that "misses" *but comes close,* and *how* close is measured in terms of the Spielraum. A still weaker case, including the extremely weak one provided by conventional null-hypothesis refutation, is when the theory is so weak it can only specify a large interval successfully (e.g., a difference will be in the upper half of the Spielraum, $M_1 - M_2 > 0$). How can we meet these desiderata for a crude index? As a first try, I suggest the following:

$S$ = Spielraum;

$I$ = interval tolerated by $T$;

$I/S$ = relative tolerance of $T$;

$In = 1 - (I/IS)$ = intolerance of $T$.

$D$ = deviation of observed value $x_0$ from edge of tolerated interval (= error);

$D/S$ = relative error;

$Cl = 1 - (D/S)$ = closeness.

Then the corroboration index $C_i$ for the particular experiment is defined as:

$$C_i = (Cl)(In),$$

that is, the product of the closeness and the intolerance. And the mean of these particular indexes (normalized in some fashion such as that to be described) over the reported experimental literature would be the cumulative corroboration $C$ of the theory.

Obviously one must supplement that index by a second number, the number of experiments. There are terrible difficulties involved in the important distinction between many replications of the same experiment and different experiments, to which I offer no solution. No mention is made of significance testing in this index, because I am not convinced that plugging it in would add anything. One would have to set up the conventional confidence belt at the edge of what the theory substantively tolerates. This is the only kind of tolerance discussed in statistics books, that due to errors of measurement and sampling in examining the statistical hypothesis $H$. The other kind of tolerance arises from the looseness, weakness, or incompleteness of $T$, and it is far more important. When we are using a correlation coefficient to test a theory, the Spielraum is the interval (–1, 1). Suppose our theory specifies a certain region of that, such as (.5, .7). Then the theory takes only a moderate risk in terms of the Spielraum. What conventional significance testing does is to focus our attention on a fuzziness at the two boundaries, that fuzziness being mainly dependent on sample size. Epistemologically, and in terms of a scientific tradition that existed in the developed sciences long before the rise of modern Fisherian statistics, that is the wrong thing to focus attention on. To include the statistician's tolerance in the corroboration index would be regressive, a shift toward strict falsification, away from verisimilitude and the "good enough" principle. This is because an *SE* probabilifies the *occurrence* of a numerical miss (i.e., a $\text{Popper}_0$ question), when what we want is *how near a miss,* as a stochastic link to verisimilitude. One could crudely state the ontological–epistemological relation thus: For "early Popper," falsification is linked to falsity, and thereby to the possibility of truth; now we link Salmonian coincidence to verisimilitude. On this emphasis, falsification does not counsel abandonment in cases of good verisimilitude.

If an index such as this, or an improved version, were applied to studying the empirical history of various scientific theories, we would begin to develop some rule-of-thumb notions about the meaning of its values for a theory's probable long-term future. That is an empirical problem for meta-theory, conceived as the rational reconstruction of history of science; more broadly, as the "science" domain of naturalized epistemology. However, I venture to suggest an a priori metric that is perhaps not devoid of merit. What is the corroboration index for an experiment that works perfectly? The observed value falls within the predicted interval, $D = 0$, and the closeness $Cl = 1$. If the theory is extremely powerful, making a very precise numerical point prediction, the allowed interval $I \simeq 0$, at least very small compared with the Spielraum, so the intolerance $In \simeq 1$. A theory that has a perfect track record in the course of 10 experiments has a cumulative index $C = \Sigma C_i/N = 1$, and we would record its track record by that index and the number of experiments thus, (1, 10).

What does the worst case look like in these terms? I don't know exactly what it means to say that a theory predicts "worse than chance," but my hunch is that if it systematically did that, it would have a funny kind of inverse verisimilitude. We would often be able to conclude something true about the state of nature from a theory that did worse than we could by flipping pennies in relation to the Spielraum. So I am going to set that case aside, and consider a theory with a dismal track record even when studied by the conventional weak form of significance testing. Our poor theory is (like most theories in soft psychology) so weak substantively that it can't predict anything stronger than a difference in a specified direction. For many situations this amounts to predicting that the observed value will be in the correct half of the Spielraum. Consider the worst scenario, in which the theory's intolerance $In = 1/2$; but despite this excessive tolerance, the theory has such poor verisimilitude that it only succeeds in predicting that direction correctly half the time (in half of the diverse experimental tests). In the basic formula multiplying the closeness, $1 - (D/S)$, by the intolerance, $I - (I/S)$, the intolerance is $1 - 1/2 = 1/2$ for a mere directional prediction. By chance this "hit" will occur half the time. For hits the deviation (error) $D_H = 0$, and the product of intolerance and closeness is

$$\begin{aligned}(In)(Cl) &= (1 - I/S)(I - D/S) \\ &= (1/2)(1 - 0) = 1/2.\end{aligned} \quad [1]$$

For "misses," where the observed value falls in the wrong half of the Spielraum, the indifference principle expects a mean untolerated point-value halfway out (middle of the residual Spielraum, $S - I$), so the expected index product for these cases is

$$(In)(Cl) = (1/2)(1 - 1/4) = 3/8. \quad [2]$$

Weighting these hit and miss values equally (hits and misses being equally probable), the expected value of the composite index for the worst case is

$$\begin{aligned}\text{Exp}(Cl) &= p_H(1/2) + p_M(3/8) \\ &= (.50)(1/2) + (.50)(3/8) \\ &= .4375 \simeq .44.\end{aligned} \quad [3]$$

If we want to normalize the cumulative index so that its range from the worst to the best case would be from 0 to 1, we would subtract this worst-case expected value from the upper ("perfect case") value = 1, and divide this difference by the constant $1 - .44 = .56$, giving the normalized cumulative index,

$$C^* = (C - .44)/.56 \quad [4]$$

which will take on value 0 for a weak theory that does no better than chance over a run of experiments, and value 1 for a strong (intolerant) theory that makes uniformly accurate point predictions. It might be just as well to apply those normalizing constants to the formula for $C_i$ itself, as computed for individual experiments (see examples in Figure 3); I have not concluded as to the merits of that, except to note that it is capable theoretically of yielding a few negative $C_i$s for "bad misses." If $C_i$ is normalized for each experiment, then the cumulative corroboration $C$ is simply the mean of the $C_i$ values (without the normalizing constants applied a second time).

Such an index would be so incomplete in appraising the theoretical situation that sole reliance on it would probably be worse than the present "informal narrative" approach to theory appraisal among working scientists. The index suffers from the defect that it conveys nothing about the total mass of experiments, nor their qualitative diversity. It is not intrinsically diachronic, although nothing prevents us from plotting its values over time. Adopting a strategy of modified auxiliaries, challenging the ceteris paribus clause, or making inroads into the peripheral postulates of the theory itself, one would also compute the index separately for the various factual domains, because the dispersion of its values over domains would presumably be related, at least loosely, to the source of the falsifications. A theory that does moderately well over all domains is a different case from one which does superlatively in some domains and fails miserably in others; and this difference provides us with guidance as to where we should begin making modifications. Despite these limitations and complications, it would be foolish to reject an index that gets at important aspects of success, such as closeness and intolerance, on the ground that it doesn't measure everything we want to take into account.

Although Popper, Lakatos, and other metatheorists hold that the ideal theory-testing situation pits competing theories against one another (probably the usual case in history of science), it is not precluded that one subjects a theory to an empirical hurdle considered solo, without a definite competitor in mind. If not falsified by the observational facts, the theory is corroborated; how strongly depends on the risk. Figure 4 illustrates several paradigm cases and is largely self-explanatory. The abscissa is an observational value, and the curves represent the net spread of corroborating values due to (a) the theory's intrinsic tolerance—a function of its incompleteness, weakness, or looseness—and (b) the statistical dispersion from errors of sampling and measurement. A theory is "weakly tested," aside from its competitor's status, if it tolerates a large region of the Spielraum. In the case of two theories, the observational value may refute both theories, or refute one and corroborate the other. Case IV is problematic because an observational value lying under $T_1$ refutes neither $T_1$ nor $T_2$, yet it seems to corroborate $T_1$ more than $T_2$ because of the latter's excessive tolerance. I believe metatheorists would disagree about that case, but I incline to think that $T_1$ is running somewhat ahead in that situation. For example, if exactly half the parents of schizophrenic probands exhibit a neurological sign (Meehl, 1962, 1989, 1990d), I would consider that corroborates a dominant-gene theory, although such a percentage is not *in*compatible with a polygenic threshold model. If the split is also one parent per pair, that would strongly corroborate the major locus conjecture; but even this finding can be adjusted ad hoc to fit a polygenic model. For obvious pictorial reasons. Figure 4 represents only Popperian "hits" and "misses," rather than the "near miss" that we count as corroborative on Salmonian coincidence grounds.

### Appraising a Theory: Function-Form Predictions

The preceding corroboration index examines the accuracy of point and interval predictions, and the chief way in which such predictions are mediated is via a specified mathematical function relating two or more observational variables. Of course the success of a theory in deriving the correct observational function is itself a strong corroborator. In advanced sciences, where one has a quasi-complete list of the elementary entities of which macro objects are composed (e.g., "corpuscularism" in the history of physics) as well as strong constraining principles (e.g., conservation laws), the theoretical derivation of a curve type may include derivation of the function parameters. In less developed sciences, or at the growing edge of the advanced sciences, the parameters may not be derivable; but having adjusted them by a suitable curve-fitting procedure, first having shown that the function chosen is a better fit than competitors, it is sometimes possible to make theory-mediated extrapolations of these parameters (or functions of them) into other experimental settings. In such cases, moving into the new experimental context serves as a more powerful corroborator because we are asking not only whether the function is a logarithm or hyperbola or straight line or whatever, but also whether the constants we plugged in, in advance of data collection, on the basis of these parameters estimated in the first experimental context, are accurate. Because the theory's ability to predict a function form is itself a corroborator, it would be helpful to have a corroboration index for that as well. Here the difficulties are greater but I think not insoluble as long as we keep in mind the modest claims appropriate for any such index in the first place.

What first occurs to one with statistical training is that it's a "goodness-of-fit" problem, so the obvious solution is something like the old correlation index, $1 - SS_R / SS_T$, the complement of the ratio of the residual variance—empirical point deviations from the curve—to the total variance. (Should the function fitted be linear, the correlation index reduces to $r^2$ = coefficient of determination.) This is easy and familiar, but quite inappropriate. The reason that it is inappropriate is that a strong theory of high verisimilitude does not necessarily rule out (a) individual differences or (b) measurement error. How large a component of total variance is contributed by these two factors will vary from one empirical domain to another and may be relatively independent of the theory's verisimilitude. (Of course, a theory that claimed to account for everything would include a prediction of individual differences. In the Utopian case it would include each individual's derivation from the best fitted function as part of what it tries to predict. This is a pipe dream for psychology and other social sciences and even for most of the biological sciences.) We do not want to fault a good theory of, say, complex human learning because we have rather unreliable measures of the output, or because there exist marked individual differences among persons; nor do we want to give too much credit to a theory in some other field where it happens that subjects differ very little and the measurement procedures are

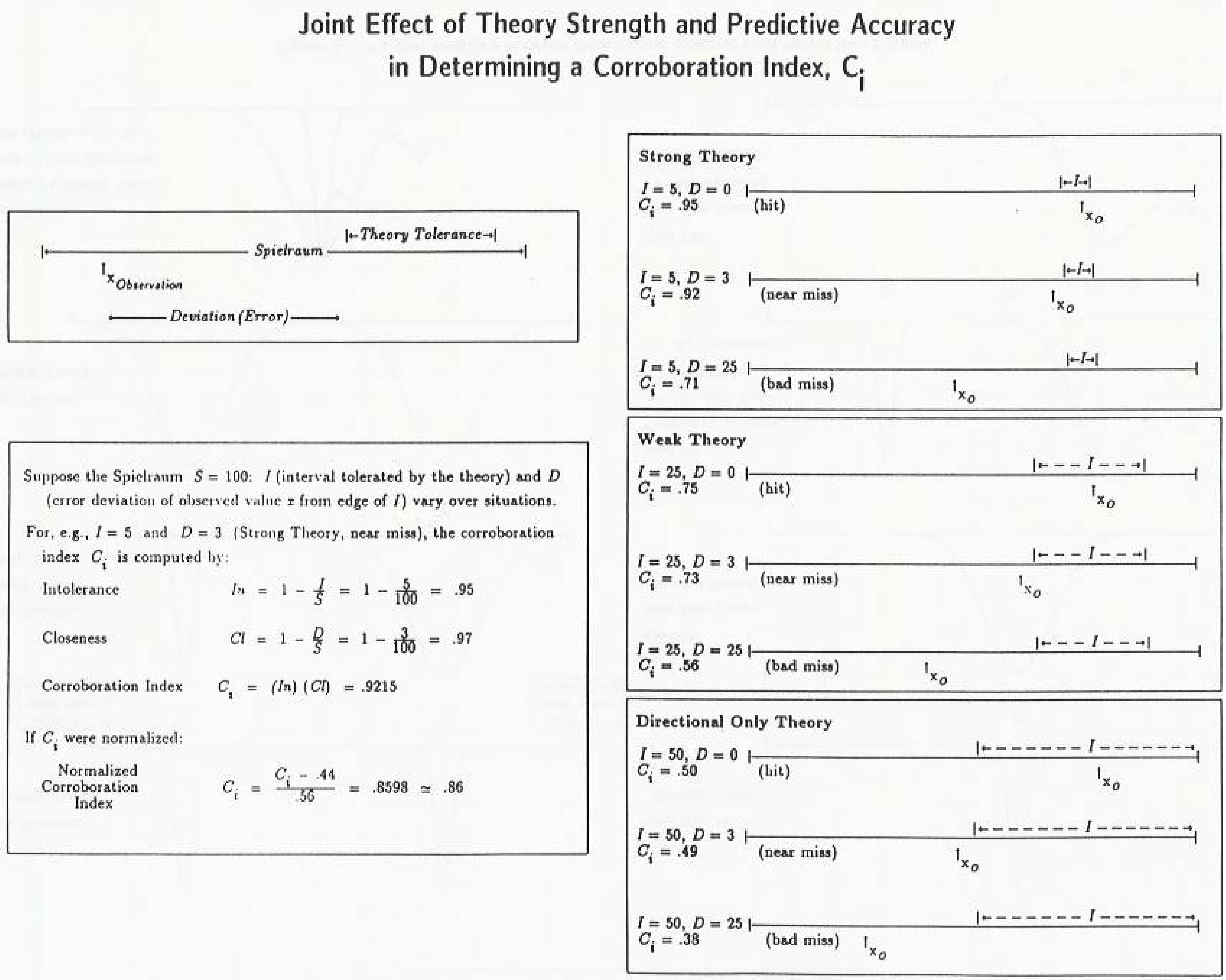


**Figure 3.** Illustration of how theory strength and predictive accuracy jointly determine a corroboration index.

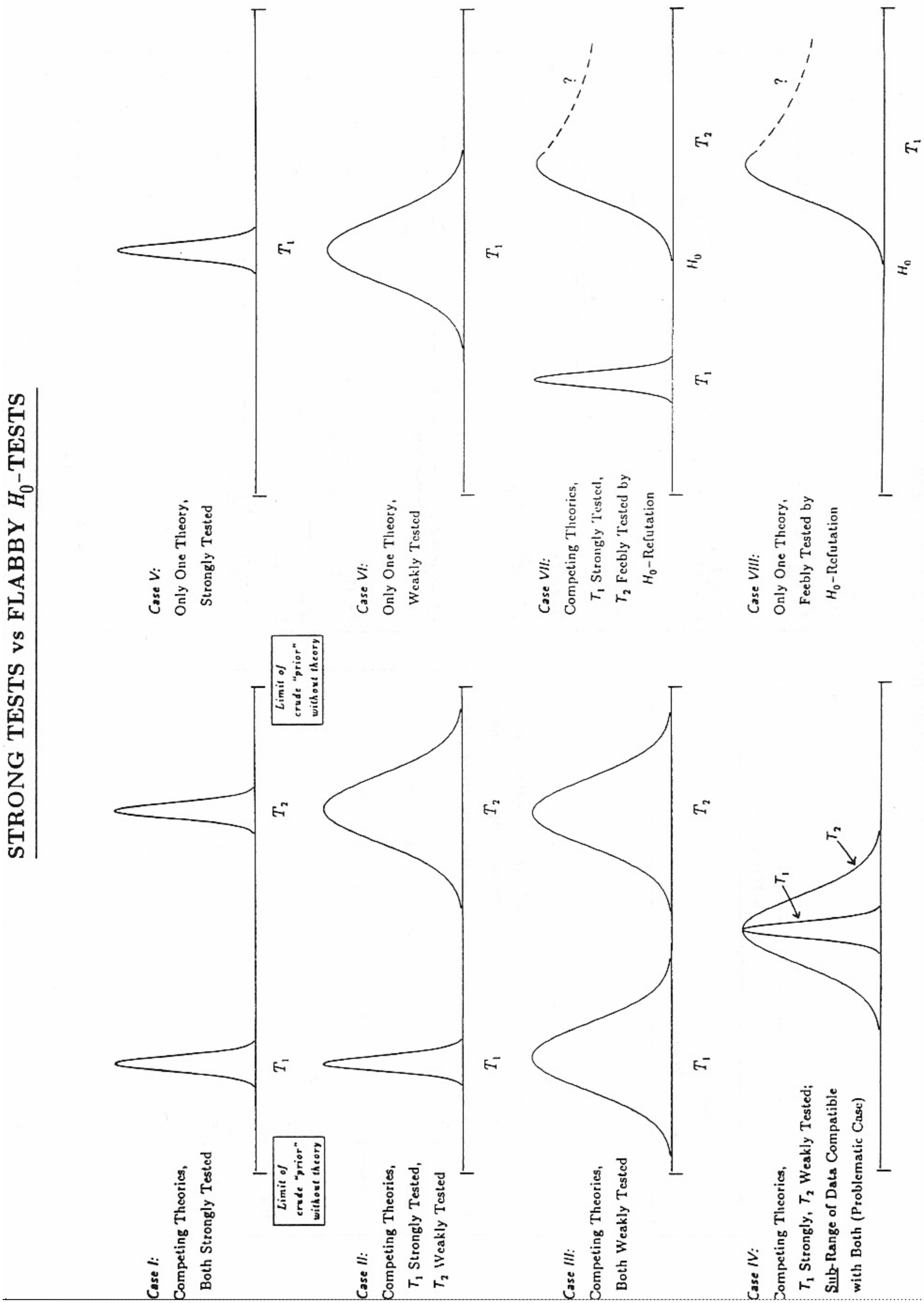


**Figure 4.** Various relations between theories and observational values they tolerate.

highly accurate, whereby the residual variance about a fitted curve remains small.

I suggest the way to deal with this is in terms of the distinction between "pure error" and "lack of fit" in regression theory (Draper & Smith, 1981). Without entering into details of the algebra, my suggestion would be this: After decomposing the total variance into the pure-error component (arising from the dispersion of individual points about the mean of an array), and the lack-of-fit component (arising from the deviations of those array means from the theoretical curve), reasoning as we do in an $F$ test that we have two independent estimates of the same variance, we estimate what the deviations of means from the theoretical curve ought to amount to on the basis of pure error. Then we compare the actual with the observed deviations of the means from the theoretical curve, thus forming an index of badness-of-fit *over and above* individual differences and measurement unreliability. The details of working out such a formula would of course depend on whether the degrees of freedom were the same in the arrays and so forth. Then, analogous to the closeness component of our corroboration index for points and intervals, we have a closeness-of-curve-type index defined as $1 - (S_m - \hat{S}_m)/\hat{S}_m$, where $S_m$ and $\hat{S}_m$ are the observed dispersion of means from the curve, and the expected dispersion of means estimated from the pure-error component, respectively. Here, as before, I wish to avoid asking the significance-test question, and for the same reasons. For example, an $F$ test may show that a parabola is a barely adequate fit, meaning that it doesn't squeak past $p = .05$. In another experiment, that same $F$ test might be at $p = .10$, considered not a significant deviation and, hence, an adequate fit. A third situation arises where the dispersion of the curve from the mean deviates hardly at all from that expected by pure error. When we are concerned with verisimilitude rather than literal truth, we do not want to lump the latter two situations together as "adequate fits" and call the first one inadequate, especially because whether we achieve a significant $F$ for a given badness-of-fit $SS_R$ depends on the power function. *We always try to minimize the influence of the power function in quantitative appraisal of verisimilitude* (Meehl, 1990c).

This crude index has to be corrected if we wish the limiting cases of excellent fit and worst scenario to behave similarly to our point or interval index, falling in the correlational interval (.00, 1.00). We do not attempt a mathematical mapping of the metric, which would be absurd to claim. But we don't want the index of closeness to take on negative values, nor do we want to give extra credit to a theory if it turns out that the dispersion of the means from the theoretical curve is *markedly less* than what pure chance predicts. In the latter case we have an "excessively good fit" that normally leads us to say not that the theory is doing beautifully, but rather that there was something hokey about the experiment! (Cf. Fisher's reanalysis of Mendel's data, indicating that he must have selected or cooked them a bit because they were closer than probability theory allows.) To avoid that undesirable consequence we may simply stipulate that if $S_m < \hat{S}_m$ we will consider the index as = 1.

What is the worst case? We want the worst scenario to be one in which the closeness index has value zero, analogously to the closeness component in the interval index. This requires that the worst case be one in which $S_m - \hat{S}_m = \hat{S}_m$—that is, that the dispersion of the means from the theoretical curve be twice what it should be as estimated from the pure-error residual. I have no intuitions about the outlandishness of such a value, but if we took that as our zero point to make the index perform properly, it would be a matter of cumulative experience to see whether we should repair it to allow a case worse than that. At first glance, it might be supposed that we could get quite a few cases worse than that by a terribly bad theory. But as I have already said, it is unclear what would be meant by negative verisimilitude, because if that arises quantitatively from indexes of one kind or another, it suggests that there is some basic truth about *what* the theory is discussing, such as the kind of entities it is postulating, and what entities are causally related to what other entities, but that the mathematical characterization of the nature of that relationship is, so to speak, "backward." I think it fruitless to consider those messy questions at this point, lacking empirical data from the history of science on the index's performance.

In defining the Spielraum of function forms, cases such as one where the theoretical curve is a parabola of high curvature convex, whereas the empirical data are well fitted by a high-curvature parabola concave, we might say the facts are almost "mirror-image opposites" in relating two variables from what the theory said they should be. This might give a badness-of-fit twice as large as that estimated from the pure-error component. However, as I discuss in a moment, this kind of thing would be prevented because two curve types of the same function form, but whose parameters lead them to be "opposite" in that graphical sense, would be treated as different functions. A parabola in the southwest and a parabola in the northeast of the graph are counted as two different function forms for Spielraum definition purposes.

Assuming we have a measure of closeness for function forms, how do we concoct a plausible measure of intolerance? We want to define a Spielraum of functions so that the prior probability of a particular function fitting the data absent the theory, or given a theory of negligible verisimilitude, will be numerified as small. That a logarithmic function, or a parabola, or a power function, or a straight line fits the data cannot constitute a Salmonian coincidence if almost all data can be fitted by a function of a given sort. (We can't get help on this from the pure mathematician, who will remind us that the number of single-valued functions $F = C^c$, the third transfinite cardinal!) We might consider as a reference class those functions that have "turned up" often enough in the various sciences and the mathematical work of pure and applied mathematicians and engineers so that it has been considered worthwhile to list them in a table of integrals. My copy *of Mathematical Tables From Handbook of Chemistry and Physics* (Hodgman, 1941) lists 322 indefinite integrals, a number that, for our purposes, is not much better than a transfiite cardinal. The point is that applying some sort of principle of indifference to a mathematician's a priori list of functions will lead to all the probabilities being less than .01, with the result that the intolerance component of our index will not be informative.

I make the following rash suggestion, which is not as crazy as it sounds when we remind ourselves that we are treating metatheory as the empirical theory of scientific theory. Theories are inscription products of the human mind, having a physical and psychological existence in Popper's Worlds I and II (I do not understand his World III, so I say nothing about it). On such a view of metatheory, we are not only allowed but required to pay attention to the empirical facts of scientific theorizing, to the scientist's cognitive dispositions. My suggestion is that for a given scientific domain, which could be broadly defined (psychology or chemistry) or more narrowly defined (the psychology of mammalian learning, or the chemistry of mammalian nutriation), we could carry out literally—by an appropriately stratified random sample of textbooks, handbooks, and research articles—a statistical study of the occurrence of the various mathematical functions. This literature survey would be diachronic, keeping track of the rate at which hitherto untallied functions appear. After a little preliminary investigation, plausible stop criteria would be set up for terminating the search, such as: "Stop when new functions are appearing at a rate less than 1 in 50 consecutive samplings, and the overall incidence of any new function, among all tokens of functions, is less than .01." From such a sampling of scientific literature, one could compile a list of functions with their relative frequency in the literature, confident that any function not found in this "function atlas" has a prior probability of less than .01 of appearing in a theory or experimental report. This finite set of functions, each occurring in empirical disciplines with nonnegligible probability, defines the Spielraum. The prior probability, "picking a function out of the function hat randomly," that it will fit a set of experimental data from the domain is then taken to be the relative frequency of that particular function in our empirical atlas.

I have not as yet made such a literature search, but I think it fairly safe to opine that better than 95% of functions that are fitted over the whole range of subdivisions of psychology would fall among the commonest 20 or fewer. Distinguishing functions as to the direction of their convexity, so that oppositely oriented hyperbolas (northwest vs. southeast) are counted as different functions for our purposes, one thinks immediately of linear functions, quadratic, cubic, quartic; polynomials above the fifth degree (these would more often be curve-fitting approximations relying on Taylor's theorem than they would be allegedly true functions); power functions (two kinds, depending on whether the exponent is greater or less than 1); exponential growth and decay functions; logistic functions; sigmoid functions (of which the Gaussian integral is a special case); Gompertz functions; hyperbolas; and certain of the common statistical functions such as gamma and beta. It doesn't take much riffling through books and articles to get quite easily to about 20 types. If they occurred with equal frequency, which of course they don't, we would have a prior probability $p = .05$ for each curve type. I dare say linear, logarithmic, exponential, and power functions would make up more than 10%, probably more like one fifth or one fourth of the functions that we run across in the life sciences.

Corresponding to the relative intolerance of the interval index, we now define the intolerance component of our function-form index simply as the empirically computed prior probability of this particular function in the given scientific domain. The "best case" (most intolerant) is taken to be one in which the prior is less than .01, that is, the function covers less than 1% of the function Spielraum. (Our crude index does not try to distinguish between a Salmonian coincidence of "chance prior probability" .008 and one of .0008, although, if that fine cutting were thought to be worthwhile, we would extend our function atlas by continuing to scan the literature until we had stable $p$ values for functions rarer than 1 %.) How do we concoct a worst case, so that the function is excessively tolerant, analogous to the weak use of significance tests for the interval index? Ignoring cases where the theory entails nothing about the relationship of the pair of observables, the weakest degree of quantification (in the earlier section on verisimilitude) is that in which we say that $x$ and $y$ are related but we characterize the relation only by the weakest statement that is semiquantitative, to wit, the first derivative is positive. When one of the observables increases, the other tends to increase also, and that is all we claim. *This is the function-form equivalent of the weak significance test when considering intervals.* One might plausibly stipulate, for purposes of an index that behaves numerically as we desire, that this prediction should have an intolerance equal to half the Spielraum. Look at it this way: If we pulled substantive theories randomly out of a theory hat, and pairs of observables randomly out of the experimental hat (as fantasized in Meehl, 1967), assuming perfect statistical power so that we don't have significance-test problems, we would expect to be "right" in stating the sign of the relation between $x$ and $y$ around half the time, in the long run. So one might say that a degree of specification of the observable relationship that does not go beyond this specificity should merit a poor intolerance component at $In = 1/2$. (I do not have any good ideas about what to do with further degrees of specification short of stating the function as being logarithmic, hyperbolic, linear, or whatever, although one might play around with the notion of similar conventions, such as, "half the time you will guess right by chance as to the sign of the second derivative," and the like.) Having defined an intolerance component and a closeness component, we again form the product, to serve as our corroboration index for function forms.

## Implausible Qualitative Predictions and Other Methods of Assessing Theories

A third kind of test that has played a crucial role in appraising scientific theories is a purely qualitative prediction which gets a lot of mileage if the qualitative event specified is unforeseeable on the basis of background knowledge and, even better, if it was taken to be intuitively implausible absent the theory. Thus, for example, some physicists dismissed the wave theory of light not only because of the prestige of Newton, but because it had been shown that, knowing roughly what the range of wavelengths had to be like, the shadow behind a shadow caster with a light source at effectively infinite distance (across a good-sized room) should produce a small spot of intense brightness in the center of the shadow. So it was strikingly corroborative of the wave theory when somebody thought he might as well try it and, lo and behold, there the bright spot was. I have no notion of how to numerify such qualitative effects, and my efforts to do it by reexpressing it quantitatively (e.g., "What is the expected size of the bright spot under those conditions?") appear highly artificial and counterintuitive.

Such suggestions concern only one major property of "good theories," namely, their ability to derive observational facts. For an empiricist (which means for any working scientist), this is doubtless the most important attribute by which one judges a theory *in the long run.* I believe that this is the basis of a final accounting of a theory's "track record," when the latter is assessed in terms of Salmon's principle or Popper's "risky test."

But I do not hold the old-fashioned logical empiricist or positivist view that this is the *only* basis on which the success of theories is appraised. The contributions of Laudan (1977) to this question of theory appraisal are of the highest importance, and I am not prepared to disagree with any of them. In psychology, I think "conceptual problems" (which he considered as important as empirical problem solving) play today, as in the past, an even greater role than in other sciences. The extent to which a theory's adequacy in problem solving of that sort would be subject to quantification by cliometric study of its documentary history is something to which I have given little thought. But I take it that at least some aspects of "conceptual fitting" involve predicting numerical values (e.g., agreement of values inferred from a reductionist view of a concept to a theory at a lower level in the pyramid of the sciences). One supposes that the same would often be true of function forms. A fair discussion of those few places where I don't quite understand Laudan, or disagree, is beyond the scope of this article. He does not deny that a major component in assessing a theory's problem-solving power is its ability to predict numerical values and function forms of observational data. If I were to offer any criticism of Laudan's book in respect to matters discussed here, it would be that (like Popper and Salmon) I attach great significance to the riskiness or "damn strange coincidence" feature of a theory's positive achievements vis-à-vis the facts, and I do not get the impression that Laudan viewed this as being so important.

### Cliometric Metatheory: Statisticizing Theory Performances

Quantifying a theory's track record, by a set of half a dozen crude indexes, might resurrect an old idea briefly mentioned by Reichenbach (1938) in defending his identity thesis concerning the probability concept against the disparity conception advocated by Carnap (1945). Prima facie, it seems odd to claim that the degree to which a diverse set of observational facts supports a theory, taken as a *probability number,* is in some deep sense a relative frequency. But Reichenbach suggested that the truth frequency of theories characterized by their possession of certain properties (both intrinsic and evidentiary?) would be the logical meaning of such degree of confirmation, on the identity conception. Because he didn't spell that out, and nobody subsequently tried to do so, the idea fell into disrepute; or perhaps one could better say it was simply ignored. On the other hand, Carnap's probability$_1$ = $p(h/e)$ = degree of confirmation, intended as a semantical concept relating hypothesis $h$ to evidence $e$ (in an ideal state-description language), was in no better shape if it came down to devising a realistic, usable numerifying algorithm for appraising theories.

Philosophers of science, when relying on a naturalized epistemology and employing history-of-science data in arguing for a rational reconstruction—with the mix of descriptive and prescriptive that *properly* characterizes metatheory on the current view—regularly do so by telling anecdotes. A reader who has not read much history of science used this way may find each philosopher's collection of anecdotes impressive, but on wider reading one doesn't know how to set them off against the opponent's favorite anecdotes. I believe this is a fundamentally defective approach to using history-of-science episodes. When Popper (1935/1959, 1983) cited an episode (e.g., the quick demise of the Bohr–Kramers–Slater quantum theory) to defend his ideas about falsification, and Feyerabend (1970) or Lakatos (1970) cited Prout's hypothesis on the other side, what do these selected episodes prove? On Popper's own view, they should all function as potential falsifiers of something, and his favorites as actual falsifiers of the opponent's view. What generalizations in empirical metascience are falsified by the two kinds of counterexamples? So far as I can make out, one kind of episode falsifies the metatheoretical statement, "No theory was ever abandoned as a result of a single clear-cut falsification of its predictions," whereas examples on the other side falsify claims that "No theory is ever successfully and fruitfully defended despite apparent falsification" and "No theory that appeared to be clearly falsified, and was as a result abandoned, has ever subsequently been resurrected in the presence of new data or new auxiliary theories." But these generalizations are not even pairwise contraries, let alone contradictories; falsifying any of them does not prove, or tend to prove, either of the others. Furthermore, it would be hard to find any scientist, or philosopher-historian of science, who has maintained any of those strong generalizations, so it seems pointless to present anecdotes involving particular episodes in the history of science to refute any of them.

Presumably philosophers of science who view metatheory as the rational reconstruction of the empirical history of science (and, therefore, as a system of formal, statistical, epistemological, and factual components) will see the enterprise as a mixture of descriptive and prescriptive statements. What they will be saying, in essence, is this: "I presuppose what most sane, informed persons will admit, that science has been, by and large, the most conspicuously successful of all human cognitive enterprises, compared with which the cognitive achievements of such disciplines as ethics, traditional political theory, 'theoretical' history, jurisprudence, aesthetics, literary criticism, theology, and metaphysics appear pretentious and often pitiable." What is it, in the way scientists go about their business, or the nature of their subject matters, that leads to this marked and indisputable superiority in knowledge claims (cf. Ziman, 1978)? If we can figure out what it is that scientists do that politicians, preachers, publicists, drama critics, and such like don't know how to do, or don't try very hard to do, we should be able to state some guidelines—not "rules" but "principles"—pieces of general advice as to how one should go about gaining reliable knowledge that brings respectable credentials with it, convinces almost all rational minds that investigate, *tends* to be cumulative, self-correcting, and technologically powerful. So we begin with a descriptive task, but we intend to conclude with some prescriptions.

In studying the history of science with this prescriptive aim in mind, one must begin by formulating the problem as a *statistical* one, not because of a psychologist's liking for statistical methods or quantification, but because the question when rightly understood is intrinsically statistical in character. No metatheoretical reconstruction of the history of science is ever going to prescribe an absolute commandment against "theoretical tenacity" (which even Popper mentions favorably in a footnote in the 1935 edition), but neither is anybody going to advise scientists, as a general policy, to stick to their guns and defend a favorite theory regardless of how degenerating the research program has become. Metatheoretical advice is like the advice to fasten your seat belt, or to buy life insurance: "This is good advice and should be followed by a rational mind." It is not refuted by the case of somebody who was strangled by a seat belt, or by the case of someone who, seeking to provide for a homemaker-spouse and five children, made the sensible move of buying a large life insurance policy, then lived to age 103, being predeceased by spouse and children, so that the death benefit went to the state. Telling such anecdotes about rare and unforeseeable events is not a rational basis to decide against fastening one's seat belt or buying life insurance. I think this is the attitude metatheorists should take in the new era of fused history and philosophy of science. Advice about a *policy* that is proffered as being "the best policy," but not "certain to win" in all cases, should be justified by showing that it increases one's *tendency* to win over what it would be if no account of this advice were taken. Why should meta-theoretical prescriptions based on the rational reconstruction of the history of science be different from practical advice of physicians, insurance counselors, psychotherapists, economists, or engineers, none of whom have the illusion that they are infallible, or that their advisory statements have the form (and intention) to be strict rules, carrying a guarantee of 100% success to those who follow them?

Smoking the cliometric opium pipe, one imagines a collection of indicators getting at distinguishable aspects of a theory's track record and a composite constructed on the basis of their statistical relationships. Suppose one had a sizable collection of minitheories going back a generation or more in the history of the science, and indexes such as the cumulative corroboration index $C$, its standard deviation over fact domains, a measure of the qualitative diversity of the fact domains, a diachronic measure of C's trend, and the like, for each minitheory. We could factor-analyze the correlation matrix of these indicators to see whether we detect a big first factor, ideally a factor accounting for nearly all the shared variance (like Spearman's $g$) for scientific theories. We could supplement this internal statistical approach by a criterion-based approach, confining ourselves initially to two sets of minitheories: (a) some that have long ago been abandoned by everyone and (b) others that have been universally accepted and appear in the textbooks as "solidly proved and not in dispute," building a linear discriminant function to predict this quasi-ultimate truth-value dichotomy. Then we ask whether the first-factor loadings of the various indicators are nearly proportional to the discriminant function weights. If so, it would be a plausible conjecture that the big statistical factor is an indicator (fallible) of a theory's verisimilitude, a stochastic thesis compatible with maintaining the distinction between verisimilitude and empirical corroboration as ontological and epistemological metaconcepts, respectively.

Scientists are bothered by this kind of thing because it sounds too mechanical, cut and dried, and hence in danger of being pseudo-objective like the kind of fake, pretentious quantification so common in the social sciences. One hesitates to substitute an equation for the wise judgment of scholars surveying the evidence in all its qualitative richness. Although I share these uneasy feelings, I suggest that they are not wholly rational, and not rational enough to be dispositive in rejecting the index idea. There is an impressive body of evidence from several disciplines indicating that informal human judgment, including that of experts and "seasoned practitioners," is not as valid as experts (and the helpless laymen who have to depend on us!) have traditionally supposed. For example:

1. It is known from studies by pathologists that the diagnostic success rate in organic medicine is much lower than the trusting patients attribute to the learned doctor (Geller, 1983; Landefeld et al., 1983; Peppard, 1949).
2. The modest reliability and validity of clinical judgment in the behavior field has been known (among sophisticated clinical psychologists) for many years, and empirical research on the relative merits of formal (statistical, mechanical, algorithmic) methods of data combination for prediction over the usual informal, impressionistic, "clinical judgment" method is remarkably consistent (Dawes, 1988; Dawes, Faust, & Meehl, 1989; Faust, 1984; Meehl, 1954, 1973b, 1986a; Sawyer, 1966; Sines, 1970).
3. In recent years, it has become a truism among philosophers and historians of science that the undergraduate stereotype of the cold, objective, superrational scientist is a myth, not warranted by the facts so far as they have been studied in a scientific way. Every informed scientist knows that there is a somewhat depressing history of resistance to scientific discoveries, that empirical findings incongruent with the received theoretical doctrines are frequently ignored or brushed aside by rather shabby ad hoc explanations, and that people pursuing novel and idiosyncratic lines of research may find it difficult to publish (Barber, 1961; Feyerabend, 1970; Fiske & Shweder, 1986; Hacking, 1988; Latour & Woolgar, 1979; Mahoney, 1976; Taton, 1957).

In recent years, there has been systematic research by cognitive psychologists and logicians into the reasoning processes of successful scientists, indicating that they frequently commit formal logical errors of a kind you would not expect sophomores to commit if they had taken an elementary logic course (Kern, Mirels, & Hinshaw, 1983). There is a growing body of research on decision making and the assessment of new evidence, both with scientists and nonscientists, which shows that there are several powerful biasing factors in the human mind, especially when large amounts of information have to be processed to arrive at a reasoned judgment (Dawes, 1988; Faust, 1984; Hogarth, 1987; Kahneman, Slovic, & Tversky, 1982; Lord, Ross, & Lepper, 1979; Nisbett & Ross, 1980). The notion that scientists reason well about the relation of theories to facts is, in addition to being flattering to us, made tempting by the obvious fact that scientific knowledge does tend to progress, to be cumulative, to bring high credentials with it, and to be amazingly powerful technologically. But that science does well when compared to other fields that make cognitive claims they cannot support (or suffer theoretical disagreements that are interminable) does not prove, or tend to prove, that scientists always reason *optimally.* That the average chemist, at least when thinking about an experiment in chemistry, "thinks better" than preachers, politicians, astrologers, soothsayers, or journalists is hardly evidence that he always thinks with beautiful clarity, rigor, and fairness. Speaking anecdotally (I have cited what I can from available quantitative data), as an amateur logician reading the arguments offered in scientific periodicals—confining myself to controversies to which I am not a party and in which I have no vested status or intellectual interest—I find that much of the reasoning is singularly shoddy. Perhaps it is due to fortunate properties of the subject matters physical and biological scientists study, and institutionalized properties of the reward system that tends (in the long run) to punish egregiously fallacious reasoning or clumsy fact collecting, that the enterprise does advance. I am as much impressed with science as anybody, and I do not suffer from the failure of nerve about science as "the best cognitive game in town" that some social scientists currently manifest; but these attitudes do not make me conclude that theory appraisal by scientists is even close to being as accurate as it might become with a little quantitative help from metatheory and naturalized epistemology.

I also take heart from the current popularity and success of the meta-analytic method in settling questions that the traditional narrative type of research summary did not succeed in settling (Glass, McGaw, & Smith, 1981; Hunter, Schmidt, & Jackson, 1982). Arguments about the instructional effect of class size (Glass, Cahen, Smith, & Filby, 1982), or the effect of psychotropic drugs (Smith, Glass, & Miller, 1980), or the efficacy of psychotherapy (Smith & Glass, 1977), had gone on for many years and did not settle these issues until the application of meta-analysis led to their being definitively answered. Meta-analysis in its received form would not, however, be the answer to our question. First, it was invented and advocated by Glass and his colleagues for evaluation research, to study the efficacy of various interventions, rather than for the testing of substantive theories; that is, its assessment aim was originally technological. Second, the basic dependent variable is effect size, the bigger the effect size the better, which is obviously not true for the testing of theories, especially strong theories which make point or narrow-interval predictions, where an effect size could err either on the high side or the low. Third, and most important, the effect size ignores the critical factor in theory testing of Popperian risk or, speaking quantitatively, of the theory's intolerance, its Salmonian coincidence. For a critique of meta-analysis as used to appraise theories, see Chow (1987).

One advantage of a composite quantitative index for theory appraisal would be to amend Reichenbach's (1938) much-criticized dichotomy between the *context of discovery* and the *context of justification* so that it would be acceptable (except to obscurantists). Although everybody agrees that Reichenbach made the distinction too easy for himself, the basic idea is surely sound; liquidating it entirely amounts to allowing what in beginning logic courses we label fallacies, such as the arguments ad personam, ad hominem, ad verecundiam, the genetic fallacy, and the like. No historian or philosopher of science would maintain that in considering the chemists' corroboration for the structure of the benzene ring we have to include, from the context of discovery, Kekulé's famous dream of the hoop snake. It is not edifying, in listening to an argument between a Freudian psychologist and one of Skinnerian persuasion, if the Freudian tells the Skinnerian that his cognitive trouble consists in not having been analyzed, or the Skinnerian reminds the Freudian how much money he spent on his analysis. So we need Reichenbach's dichotomy, but we have to clean it up. One way to do this is to think in terms of metatheory as the rational reconstruction of the history of science, in which the prescriptive features of metatheory are derived by a combination of the descriptive features with some a priori components from logic, probability theory, and pure epistemology (cf. Meehl, 1984). I say again, we start with the common-sense observation that science is, by and large, a remarkably successful enterprise in finding out the way things work. Granting that, we would like to know what it is that scientists do better than others who engage in cognitive; enterprises that are not attended with the scientists' conspicuous success in solving their problems. Research strategies and methods of theory appraisal that could be "validated" by a cliometric approach to the history of science would then be formulated as rules of thumb, guidelines, and pieces of friendly advice, including the advice that a few brilliant mavericks should, from time to time, deviate from the guidelines.

One can even imagine a composite index for theory appraisal coming to have some pragmatic value—first, for the individual scientist or laboratory in adopting research strategy and tactics; second, for funding agencies which have to make such appraisals willy-nilly when resources are limited; and even conceivably for academic departments when assigning priorities in personnel recruitment. The state of various theories and research programs is currently being appraised at all these levels, unavoidably; so objections to the index idea cannot fairly be, "Who dares to appraise?" Rather, objections must be based on the belief that an informal, cryptoquantitative appraisal is better than a formal, explicitly quantitative one. I do not think this belief can be sustained either from the armchair or based on our available empirical evidence about human cognitive processes.

## Is It *Ever* Correct to Use Null-Hypothesis Significance Tests?

Of course it is. I do not say significance testing is *never* appropriate or helpful; there are several contexts in which I would incline to criticize a researcher who failed to test for significance. The first involves technological problems, where we are not (primarily) interested in examining the verisimilitude of an explanatory theory but rather in evaluating a technique (tool, procedure, action) aimed at some pragmatic end. If we compare two antidepressants in a psychopharmacological study, and one drug helps 7% more patients than the other, we want to know whether that 7% can be plausibly attributed to "chance" before advising practitioners or drug companies as to the merits. However, even here I would urge the superiority of setting up a confidence belt, which would give us additional information as to the size of a difference with specified levels of confidence. There may even be some situations where the pragmatic context is such that we ought to rely on an observed difference whatever its significance level (assuming costs and adverse side effects to be equal). As was pointed out many years ago (Simon, 1945), the best estimate of a mean, the best estimate of a proportion, and *the best estimate of a difference between two means or proportions is the observed one, quite apart from significance testing.* So that if sulfadiazene produced grave kidney pathology in 7% more children with strep throat than penicillin did, but the sample was so small that this difference was not statistically significant (even, say, at the 25% level of confidence), utility theory might justify, pending more data with large samples having higher statistical power, preferring penicillin in the meantime.

A second context is that in which there is essentially no difference between the content of the substantive theory $T$ and the counternull statistical hypothesis $H^*$, so that refuting $H_0$ (thereby corroborating $H^*$) is equivalent to corroborating $T$. It is this fact of a negligible "semantic distance" between the content of $T$ and $H^*$ that leads to the legitimate reliance on significance testing in such fields as agronomy, where the difference between the statement "those plots that were fertilized yielded more corn" and the statement "it helps to grow corn if you fertilize it" is of no consequence except in a seminar on Hume (Meehl, 1978, 1990c). When I was a rat psychologist, I unabashedly employed significance testing in latent-learning experiments; looking back I see no reason to fault myself for having done so in the light of my present methodological views. Although Tolman's cognitive theory was not sufficiently strong to make quantitative predictions, or even predictions of function forms, it did insist that the rat could learn "about the maze" or "how to get somewhere" or "where something can be found" in other ways than by strengthening a stimulus–response (SR) connection by contingent reinforcement. By contrast, Hull's theory, or other SR drive-reduction or reinforcement theories, implied that any learning the rat did was either (a) the acquisition of reinforcing power by a stimulus or (b) the strengthening of an SR connection. There were, of course, some difficult problems about the auxiliaries and ceteris paribus clauses; but setting them aside, these two competing theories of maze learning involve the assertion and the denial that under certain conditions *something,* as contrasted with *nothing,* would be learned. When that difference between learning something and nothing is translated into comparison of the experimental and control group, we have a case similar to that of agronomy (although admittedly not quite as clean); and a showing that the rat did learn something when it was not manifesting evidence of a strengthened SR connection, or when it was not being rewarded at the end of a behavior sequence, was almost equivalent to showing that cognitive theory was correct and SR reinforcement theory was wrong.

Third, even in the context of discovery (Reichenbach, 1938) there do occur rational (critical, evaluative) components, considerations that normally we assign to the context of justification. Adoption of a research program, or preference for one type of apparatus rather than another to study a phenomenon such as latent learning, is not done by the scientist whimsically or intuitively, but with rational considerations in mind. Investigator B reads an article by investigator A claiming a certain effect was obtained. Before deciding whether to try replicating this, or modifying the experiment to get outcomes different from those A reported, it is rational for B to inquire whether A's result could easily have arisen "by chance alone." This is close to asking whether the phenomenon is reproducible, and it is more likely to be reproducible if A found it to be statistically significant than if not. Yet even this case highlights a basic point made by Skinner years ago in his classic 1938 volume where he felt under some pressure to explain why he had not done any significance tests. A scientific study amounts essentially to a "recipe," telling other cooks how to prepare the same kind of cake the recipe writer did. If other competent cooks can't bake the same kind of cake following the recipe, then there is something wrong with the recipe as described by the first cook. If they can, then, the recipe is all right, and has probative value for the theory. It is hard to avoid the thrust of the claim: *If I describe my study so that you can replicate my results, and enough of you do so, it doesn't matter whether any of us did a significance test; whereas if I describe my study in such a way that the rest of you cannot duplicate my results, others will not believe me, or use my findings to corroborate or refute a theory, even if I did reach statistical significance.* So if my work is replicable, the significance test is unnecessary; if my work is not replicable, the significance test is useless. I have never heard a satisfactory reply to that powerful argument.

It is interesting that the grip of the received research tradition is so strong that some insist on significance tests in settings where data are so clear and the reproducibility so good that scientists in other fields would not bother with statistics. I am told by reliable witnesses that there are accredited psychology departments in which the faculty is so hidebound by Fisherian design that a student's dissertation will not be accepted unless it includes an analysis of variance, studying higher-order interactions, using Greco-Latin squares, and the like. Such a department would presumably have refused to grant a doctorate to most of the great scientists in physics, chemistry, astronomy, geology, medicine, or biology prior to 1925! I think this is absurd. My late colleague Kenneth McCorquodale wrote his doctoral dissertation on data from air crew pilots in the Navy during World War II; the problem was the blindfolded subject's ability to discriminate "tilt" and "turn" from proprioceptive and vestibular cues alone. The data were orderly, consistent, and the trends powerful; the graphs of verbal reports as a function of degree of tilt and turn showed quite clearly how the discriminations were working. Despite this clear-cut order, an educational psychologist on his examining committee observed, "These are certainly beautiful curves you got …"and then added almost wistfully, "but, couldn't you somewhere work in a few $t$ tests?" That is pathetic.

In either a theoretical or technological context, replicability (preferably by different workers) is more important than statistical significance. Suppose a single investigator reports a difference between two drugs favoring A over B, significant at the $p = .05$ level. Would we prefer, as clinicians, to have this information rather than learning that four different laboratories (none of which reported a significance test) all found drug A superior, yielding a sign test at $p = .06$? I think not. The improbability of the total evidence being "due to chance" is roughly the same, although the four-study situation fails to squeak by the magic .05 level. The methodological and epistemological (some would say "sociological") merits of four labs agreeing are too well known to require exposition here, and they are far more important than the difference between .05 and .06, or even a larger discrepancy than that one.

## Conclusion

I have tried to provide a reformulation of Serlin and Lapsley's (1985) "good enough" principle that preserves the Popperian emphasis on strong corroboration. Accepting their criticism of my overly strict Popperian formulations, and moving from Popper to Lakatos as a metatheoretical guide, we ask not, "Is the theory literally true?" but instead, "Does the theory have sufficient verisimilitude to warrant our continuing to test it and amend it?" This revised appraisal in terms of verisimilitude rather than strict truth leads to adopting a strategy of Lakatosian defense by strategic retreat, provided the ad hockery is "honest" at all stages (i.e., *not* ad hoc in any of Lakatos's three senses). The warrant for conducting a Lakatosian defense is the theory's track record. A good track record consists of successful and almost-successful risky predictions, of "hits" and "near misses" for point or interval predictions of low tolerance, and predictions of function forms. It is crucial in my argument that this low tolerance is not best judged by traditional significance testing, whether of the strong or weak kind, or even by confidence-interval estimation, but by comparing the theory's intolerance, and the nearness of the "miss" when there is a miss, with a reasonable a priori range of possible values, the antecedent Spielraum. *Whether my specific proposals for quantitative indexes of corroboration are acceptable is not the main point.* The big qualitative point is Salmon's principle. It would be unfortunate if accepting some form of the good-enough principle that still emphasizes significance testing, especially of the weak kind, the mere refutation of $H_0$, should blunt the attack on that tradition by Bakan (1966), Carver (1978), Chow (1988), Lykken (1968), Meehl (1967, 1978, 1990c), Rozeboom (1960), and others (see Morrison & Henkel, 1970).

I hope my acceptance of Serlin and Lapsley's criticism of too-strong falsificationism is not taken as recanting what I have written about feeble significance testing of weak theories, nor the distinction between the strong and weak use of significance testing in physics and psychology, respectively. Let me say as loudly and clearly as possible that what we critics of weak significance testing are advocating is not some sort of minor statistical refinement (e.g., one-tailed or two-tailed test? unbiased or maximum likelihood statistics? pooling higher order uninterpretable and marginal interactions into the residual?). It is not a reform of significance testing as currently practiced in soft psychology. We are making a more heretical point than any of these: *We are attacking the whole tradition of null-hypothesis refutation as a way of appraising theories.* The argument is intended to be revolutionary, not reformist. So, although I cheerfully confess error in being such a strict Popperian 20 years ago and admit incompleteness in assimilating Lakatos a decade ago, I emphasize in closing that one gets to Lakatos via Popper. Most psychologists using conventional $H_0$-refutation in appraising the weak theories of soft psychology have not reached the stage of $\text{Popper}_0$ and are living in a fantasy world of "testing" weak theories by feeble methods.

## Note

Paul E. Meehl, Department of Psychology, N218 Elliott Hall, University of Minnesota, 75 East River Road, Minneapolis, MN 55455.

## 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.

Aronson, E., Willerman, B., & Floyd, J. (1966). The effect of a pratfall on increasing interpersonal attractiveness. *Psychonomic Science, 4,* 227-228.

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66,* 423-437. (Reprinted in D. E. Morrison & R. E. Henkel [Eds.], *The significance test controversy* [pp. 231-251]. Chicago: Aldine, 1970)

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

Betz, N. E. (Ed.). (1986). The *g* factor in employment [Special issue]. *Journal of Vocational Behavior,* 29(3).

Brink, C., & Heidema, J. (1987). A verisimilar ordering of theories phrased in propositional language. *British Journal for Philosophy of Science, 38,* 533-549.

Carnap, R. (1936-1937). Testability and meaning. *Philosophy of Science, 3* 420-471; *4,* 2-40. (Reprinted with corrigenda and additional bibliography. New Haven, CT: Yale University Graduate Philosophy Club. 1950; and in H. Feigl & M. Broadbeck [Eds.], *Readings in the philosophy of science* [pp. 47-92]. New York: Appleton-Century-Crofts, 1953)

Carnap, R. (1945). The two concepts of probability. *Philosophy and Phenomenological Research, 5,* 513-532. (Reprinted in H. Feigl & M. Broadbeck [Eds.], *Readings in the philosophy of science* [pp. 438-455]. New York: Appleton-Century-Crofts, 1953)

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

Cartwright, N. (1983). *How the laws of physics lie.* New York: Oxford University Press.

Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review, 48,* 378-399.

Chow, S. L. (1987). Meta-analysis of pragmatic and theoretical research: A critique. *Journal of Psychology, 121,* 259-271.

Chow, S. L. (1988). Significance test or effect size? *Psychological Bulletin, 103,* 105-110.

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* [pp. 3-31]. Minneapolis: University of Minnesota Press, 1973)

Dar, R. (1987). Another look at Meehl, Lakatos, and the scientific practices of psychologists. *American Psychologist, 42.* 145-151.

Dawes, R. M. (1988). *Rational choice in an uncertain world.* Chicago: Harcourt Brace Jovanovich.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus statistical prediction of human outcomes. *Science, 243,* 1668-1674.

Draper, N., & Smith, H. (1981). *Applied regression analysis* (2nd ed.). New York: Wiley.

Dworkin, R. M. (1967). The model of rules. *University of Chicago Law Review, 35,* 14-46.

Eisberg, R. M. (1961). *Fundamentals of modern physics.* New York: Wiley.

Evans, C., & McConnell, T. R. (1941). A new measure of introversion–extroversion. *Journal of Psychology, 12,* 111-124.

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

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* (pp. 17-130). Minneapolis: University of Minnesota Press.

Fisher, R. A. (1925). *Statistical methods for research workers.* London: Oliver & Boyd.

Fisher, R. A. (1937). *The design of experiments* (2nd ed.). London: Oliver & Boyd.

Fiske, D. W., & Shweder, R. A. (1986). *Metatheory in social science.* Chicago: University of Chicago Press.

Freud, S. (1957). On the history of the psycho-analytic movement. In J. Strachey (Ed. & Trans.), *The standard edition of the complete psychological works of Sigmund Freud* (Vol. 14, pp. 7-66). London: Hogarth. (Original work published 1914)

Geller, S. A. (1983, March). Autopsy. *Scientific American, 248,* 124-136.

Giere, R. N. (1984). *Understanding scientific reasoning.* New York: Holt. Rinehart & Winston.

Giere, R. N. (1988). *Explaining science: A cognitive approach.* Chicago: University of Chicago Press.

Glass, G. V., Cahen, L. S., Smith, M. L., & Filby, N. K. (1982). *School class size: Research and policy.* Beverly Hills, CA: Sage.

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 University Press.

Golden, R., & Meehl, P. E. (1978). Testing a single dominant gene theory without an accepted criterion variable. *Annals of Human Genetics London, 41,* 507-514.

Goldstick, D., & O'Neill, B. (1988). "Truer." *Philosophy of Science, 55,* 583-597.

Gough, H. G. (1987). *CPl, Administrator's guide.* Palo Alto, CA: Consulting Psychologists Press.

Hacking, I. (1988). The participant irrealist at large in the laboratory. *British Journal for the Philosophy of Science, 39,* 277-294.

Harman, H. H. (1960). *Modern factor analysis.* Chicago: University of Chicago Press.

Hedges, L. V. (1987). How hard is hard science, how soft is soft science? *American Psychologist, 42,* 443-455.

Hodgman, C. D. (Compiler). (1941). *Mathematical tables from handbook of chemistry and physics* (7th ed.). Cleveland, OH: Chemical Rubber Publishing Co.

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

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

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). *Meta-analvsis: Cumulating research findings across studies.* Beverly Hills, CA: Sage.

Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgments under uncertainty: Heuristics and biases.* Cambridge, England: Cambridge University 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.

Kuhn., T. S. (1970). The structure of scientific revolutions (2nd ed.; Vol. 2, No. 2 of *International Encyclopedia of Unified Science*). Chicago: University of Chicago Press.

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

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

Lakatos, I. (1971). History of science and its rational reconstructions. In R. C. Buck & R. S. Cohen (Eds.), *P. S. A.* (1970 Boston Studies in the Philosophy of Science, Vol. 8, pp. 91-135). Dordrecht, Netherlands: Reidel. (Reprinted in J. Worrall & G. Currie [Eds.], *Imre Lakatos: Philosophical papers. Vol. I: The methodology of scientific research programmes* [pp. 102-138]. New York: Cambridge University Press. 1978)

Landefeld, C. S., Chren. M., Myers, A., Geller, R., Robbins, S., & Goldman, L. (1983). Diagnostic yield of the autopsy in a university hospital and a community hospital. *New England Journal of Medicine, 318.* 1249-1254.

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: University of California Press.

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

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.

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin, 70*, 151-159. (Reprinted in D. E. Morrison & R. E. Henkel [Eds.]. *The significance test controversy* [pp. 267-279], Chicago: Aldine, 1970)

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* (pp. 177-266). New York: Appleton-Century-Crofts.

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

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

Maxwell, G. (1970). Structural realism and the meaning of theoretical terms. In M. Radner & S. Winokur (Eds.), *Minnesota studies in the philosophy of science: Vol. 4. Analyses of theories and methods of physics and psychology* (pp. 181-192). Minneapolis: University of Minnesota Press.

McClosky, H., & Meehl, P. E. (in preparation). *Ideologies in conflict.*

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

Meehl, P. E. (1962). Schizotaxia, schizotypy, schizophrenia. *American Psychologist, 17,* 827-838. (Also available in P. E. Meehl, *Psychodiagnosis: Selected papers* [pp. 135-155]. Minneapolis: University of Minnesota Press, 1973)

Meehl, P. E. (1964). *Minnesota-Ford Project genotypic personality items* [IBM and compatibles machine-readable data file]. (Available, with the Minnesota-Ford pool of phenotypic personality items, from P. E. Meehl, Department of Psychology, N218 Elliott Hall, University of Minnesota. 75 East River Road, Minneapolis, MN 55455)

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34,* 103-115. (Also available in D. E. Morrison & R. E. Henkel [Eds.]. *The significance test controversy* [pp. 252-266]. Chicago: Aldine, 1970)

Meehl, P. E. (1971). High school yearbooks: A reply to Schwarz. *Journal of Abnormal Psychology, 77,* 143-148. (Also available in P. E. Meehl. *Psychodiagnosis: Selected papers* [pp. 174-181]. Minneapolis: University of Minnesota Press, 1973)

Meehl, P. E. (1972). Specific genetic etiology, psychodynamics and therapeutic nihilism. *International Journal of Mental Health, 1,* 10-27. (Also available in P. E. Meehl, *Psychodiagnosis: Selected papers* [pp. 182-199]. Minneapolis: University of Minnesota Press, 1973)

Meehl, P. E. (1973a). MAXCOV-HITMAX: A taxonomic search method for loose genetic syndromes. In P. E. Meehl, *Psychodiagnosis: Selected papers* (pp. 200-224). Minneapolis: University of Minnesota Press.

Meehl, P. E. (1973b). *Psychodiagnosis: Selected papers.* Minneapolis: University of Minnesota Press.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46,* 806-834.

Meehl, P. E. (1983a). Consistency tests in estimating the completeness of the fossil record: A neo-Popperian approach to statistical paleontology. In J. Earman (Ed.), *Minnesota studies in the philosophy of science: Vol. X. Testing scientific theories* (pp. 413-473). Minneapolis: University of Minnesota Press.

Meehl, P. E. (1983b). 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* (pp. 349-411). Minneapolis: University of Minnesota Press.

Meehl. P. E. (1984). Foreword. In D. Faust. *The limits of scientific reasoning* (pp. xi-xxiv). Minneapolis: University of Minnesota Press.

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

Meehl, P. E. (1986b). Diagnostic taxa as open concepts: Metatheoretical and statistical questions about reliability and construct validity in the grand strategy of nosological revision. In T. Millon & G. L. Klerman (Eds.). *Contemporary directions in psychopathology* (pp. 215-231). New York: Guilford.

Meehl, P. E. (1989). Schizotaxia revisited. *Archives of General Psychiatry, 46,* 935-944.

Meehl, P. E. (1990a). *Corroboration and verisimilitude: Against Lakatos' "sheer leap of faith"* (Working paper). Minneapolis: Minnesota Center for Philosophy of Science.

Meehl, P. E. (1990b). Schizotaxia as an open concept. In A. I. Rabin, R. Zucker, R. Emmons, & S. Frank (Eds.). *Studying persons and lives* (pp. 248-303). New York: Springer.

Meehl, P. E. (1990c). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports, 66,* 195-244. In R. E. Snow & D. Wiley [Eds.], *Improving inquiry in social science: A volume in honor of Lee J. Cronbach* (pp. 13-59). Hillsdale. NJ: Lawrence Erlbaum Associates, Inc.)

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

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

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.

Meehl, P. E., Schofield, W., Glueck, B. C., Studdiford, W. B., Hastings, D. W., Hathaway, S. R., & Clyde, D. J. (1962). *Minnesota-Ford pool of phenotypic personality items* (August 1962 ed.). Minneapolis: University of Minnesota.

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

Neurath, O. (1932–1933). Protokollsätze. *Erkenntnis, 3.* (Trans. as "Protocol sentences" by F. Schick in A. J. Ayer [Ed.], *Logical positivism* [pp. 201-204]. Glencoe, IL: Free Press, 1959)

Nisbett, R. E., & Ross, L. (1980). *Human inference: Strategies and shortcomings of human judgment.* Englewood Cliffs, NJ: Prentice-Hall.

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

O'Hear. A. (1980). *Karl Popper.* Boston: Routledge & Kegan Paul.

Pap, A. (1953). Reduction-sentences and open concepts. *Methodos, 5,* 3-30.

Pap, A. (1958). *Semantics and necessary truth.* New Haven, CT: Yale University Press.

Pap, A. (1962). *An introduction to the philosophy of science.* New York: Free Press.

Peppard, T. A. (1949). Mistakes in diagnosis. *Minnesota Medicine, 32,* 510-511.

Perrin, J. B. (1916). *Atoms* (D. L. Hammick, Trans.). New York: Van Nostrand. (Original work published 1913)

Popper, K. R. (1959). *The logic of scientific discovery.* New York: Basic. (Original work published 1935)

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

Popper.K. R. (1983). *Postscript to the logic of scientific discovery: Vol. I. Realism and the aim of science*, W. W. Bartley III (Ed.). Totowa, NJ: Rowman & Littlefield.

Rawls, J. (1971). *A theory of justice.* Cambridge, MA: Harvard University Press, Belknap Press.

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

Rozeboom, W. W. (1960). The fallacy of the null hypothesis significance test. *Psychological Bulletin, 57.* 416-428. (Reprinted in D. E. Morrison & R. E. Henkel [Eds.], *The significance test controversy* [pp. 216-230]. Chicago: Aldine, 1970)

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

Sawyer, J. (1966). Measurement *and* prediction, clinical *and* statistical. *Psychological Bulletin, 66,* 178-200.

Schilpp, P. A. (Ed.). (1974). *The philosophy of Karl Popper.* LaSalle. IL: Open Court.

Serlin, R. C., & Lapsley, D. K. (1985). Rationality in psychological research: The good enough principle. *American Psychologist, 40,* 73-83.

Simon, H. A. (1945). Statistical tests as a basis for "yes-no" choices. *Journal of the American Statistical Association, 40,* 80-84.

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.

Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist, 32,* 752-760.

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

Suppe, F. (Ed.). (1977). *The structure of scientific theories* (2nd ed.). Urbana: University of Illinois Press.

Taton, R. (1957). *Reason and chance in scientific discovery.* London: Hutchinson.

Thurstone, L. L. (1938). *Primary mental abilities.* Chicago: University of Chicago Press.

Thurstone, L. L., & Thurstone, T. G. (1941). *Factorial studies of intelligence.* Chicago: University of Chicago Press.

Worrall, J., & Currie, G. (Eds.). (1978a). *Imre Lakatos: Philosophical papers. Vol. 1: The methodology of scientific research programmes.* New York: Cambridge University Press.

Worrall, J., & Currie, G. (Eds.). (1978b). *Imre Lakatos: Philosophical papers. Vol. 2: Mathematics, science and epistemolog\.* New York: Cambridge University Press.

Wren, P. C. (1925). *Beau Geste.* New York: Frederick A. Stokes.

Ziman, J. (1978). *Reliable knowledge.* New York: Cambridge University Press.

pdf by ljy May 2003

### Commentaries were provided by:

**Campbell, Donald T.** (1990) The Meehlian corroboration-verisimilitude theory of science. *Psychological Inquiry, 1*, 142-147.

**Chow, Siu L.** (1990) In defense of Popperian falsification. *Psychological Inquiry, 1*, 147-149.

**Dar, Reuven** (1990) Theory corroboration and football: Measuring progress. *Psychological Inquiry, 1*, 149-151.

**Fiske, Donald W.** (1990) Judging results and theories. *Psychological Inquiry, 1*, 151-152.

**Humphreys, Lloyd G.** (1990) View of a supportive empiricist. *Psychological Inquiry, 1*, 153-155.

**Kimble, Gregory** (1990) A grivial disagreement? *Psychological Inquiry, 1*, 156-157.

**Kitcher, Philip** (1990) The complete falsifier. *Psychological Inquiry, 1*, 158-159.

**Kukla, Andre** (1990) Clinical versus statistical theory appraisal. *Psychological Inquiry, 1*, 160-161.

**Maxwell, Scott E., & Howard, George S.** (1990) Thoughts on Meehls vision of psychological research for the future. *Psychological Inquiry, 1*, 162-164.

**McMullin, Ernan** (1990) The Meehlian corroboration-verisimilitude theory of science. *Psychological Inquiry, 1*, 164-166.

**Rorer, Leonard G.** (1990) The limits of knowledge: Bayesian pragmatism versus a Lakatosian defense. *Psychological Inquiry, 1*, 167-168.

**Serlin, Ronald C., & Lapsley, Daniel K.** (1990) Meehl on theory appraisal. *Psychological Inquiry, 1*, 169-172

# AUTHOR'S RESPONSE

I am grateful to those who made comments on my article, for their laudatory remarks, and for making me clarify and rethink the ideas. Whomever readers agree with, they will profit immensely from the exchange. First I respond to some specific points made by each of the commentators (in alphabetic order); then I continue with a more focused discussion of my corroboration index and verisimilitude, and statisticizing in general.

## Responses

### Campbell

Cronbach and I (1955) were still too much logical positivists in our discussion of the nomological net, although I believe our emphasis on *bootstrap effect, open concepts,* and *early stages* was liberating. One should remember that the positivists themselves had made significant advances in that direction, as Pap (1953) pointed out in his classic article. If forced to assign a date to the demise of Vienna positivism, I would say 1950, the year Feigl, who *invented* the phrase "logical positivism" and co-authored the first article in English introducing it to us (Blumberg & Feigl, 1931), published his neglected article on existential hypotheses (Feigl, 1950a). Clustered around that date are MacCorquodale and Meehl (1948); Waismann (1945); Carnap (1936–1937, 1956); articles by Carnap (1950), Feigl (1950b), and Hempel (1950) in the *Revue Internationale de Philosophie;* Pap (1953); and Cronbach and Meehl (1955). As to permitting discretionary judgments, an index such as $C_i$ aims to *aid* and *contain* them, and I still hold that some observational facts are not theory laden (e.g., "Rat 3 turned right"; cf. Meehl, 1983, pp. 389 ff.). I am not sure that I want to emancipate myself further from the positivist framework, and, although I admit I am offering a psychology of science, it is intended to include prescriptive, normative components. I do not think I exaggerate the role of theory (also suggested by Fiske), hardly possible for a Minnesota PhD with undergraduate advisor D. G. Paterson and graduate advisor S. R. Hathaway! My early work on validation of the Minnesota Multiphasic Personality Inventory (MMPI) was only minimally "theoretical," and, as a practicing therapist, I believe strongly in "exploratory" and "refined folk-observational" knowledge. The article, however, was *about* theory testing, which perhaps leads to a wrong impression of my overall emphasis. As to explicating Popper's emphasis on prediction over ad hoc convergence (which Carnap and others never accepted), see Meehl (1990). Campbell may be correct that I owe less to Lakatos than I thought I did, and I do not take much from his diachronic emphasis, or from some other aspects of his approach.

### Chow

Focusing on soft psychology does tend to make one less a Popperian falsificationist than does working in a strong experimental domain. But Lakatosian defense also occurs, often appropriately, in the latter, "at reasoned discretion," but not dogmatically. No one today knows how best to do that, and my article offers no suggestions. That some auxiliaries have been independently corroborated so strongly that challenging them is poor tactics, I take for granted and should have mentioned explicitly.

### Dar

Dar's comments were mainly concerning my proposed corroboration index and are addressed in my subsequent discussion.

### Fiske

My term "the theorist" individualizes the scientist, but of course I agree with Fiske about theorists functioning in a social context. Yet "the scientific club" is composed of members, and any group consensus (deciding what can now go in textbooks, or in an encyclopedia) is based on what individual scientists have concluded. Broad agreement as to position (Freudian, Skinnerian) allows considerable leeway, fortunately, for cooperative research as well as applications. Whether successful predictions from "my version" of a theory put money in the bank for your version would depend on making our differences explicit. If we share a core postulate $P_1$ and the derivation chain to the predicted fact involves $P_1$, did my postulate $P_2$ play an essential role that your $P'_2$ cannot play? This presents knotty problems for the logician, but see Meehl (1990). I am gratified that Fiske sees clearly that I am "advocating an approach, not a technique." Had I said that in those terms, my other critics would have been saved some trouble.

### Humphreys

I agree that methodological worries are usually the concern of immature science, although advanced sciences do often experience these stomachaches in Kuhnian crises. In quantum mechanics, there have been persisting "philosophical" worries for more than half a century. I, too, think Kuhn's impact on the soft areas of psychology is unhealthy. That Humphreys arrives at views similar to mine without reading philosophy is reassuring to me. (It might also suggest that Humphreys has a natural talent for philosophy of science whether he likes it or not!) His discussion of hypothesis testing is a nice scientist's filling out, from technical considerations in statistics, of my position. Like him, I am puzzled by the psychologists' neglect of confidence intervals for significance tests, because the formalism and numerical values are identical for most applications. My adoption of Lykken's "crud factor" terminology (in his 1968 article, he labeled it "ambient noise level," but for years we have regularly said "crud factor" in conversation) may be unfortunate, and *systemic noise* would be better. My colleague Auke Tellegen complained of this after reading the manuscript, and I should have taken his advice. It even misled Kitcher into thinking I meant statistical error, although my text does say explicitly that it denotes real (stable, replicable) correlations due to all the causal influences, known and unknown, at work in a domain. As to Humphreys's preference for having good data before embarking on theories, here is one of several places that I am not strongly Popperian, as I agree with Humphreys. But in agreeing I mean *good* data, not necessarily *a lot* of good data. Small amounts of good data, especially if qualitatively diverse, suffice to warrant embarking on bold conjectures.

**Kimble**

Kimble agrees with me almost entirely and provides a nice restatement of my general position. As to what he calls his "quibble," I cannot respond to it, because it presupposes rejection of my distinction between the weak and strong use of significance tests without his saying why that distinction is invalid. So what, given that unexplained threshold difference, can I say in rejoinder? I agree that too often psychologists fiddle with theoretical adjustments instead of making sure the discordant factual finding replicates. Lack of replication is one of the worst defects of social science, to which my article perhaps gave insufficient attention (because I assumed we all know about it and deplore it). Like Kimble, I hope no one takes my critique of $H_0$-refutation as suggesting we "abandon statistical thinking." One who became famous overnight by Meehl (1954) is hardly likely to be "against statistics," and, of course, index $C_i$—whatever its defects—is inherently statistical, in the broad sense.

**Kitcher**

I agree with Kitcher about "the overall epistemic goodness of the bodies of belief that would result from various modifications." Whether this consideration necessarily renders my $C_\mathrm{i}$ index too atomistic I do not know. In the article, I did not say when (whether?) one should recompute such an index for the new conjunction $T \,.\, A,$ because I simply had not thought about it. It will, I fear, require more thought and discussion than my deadline permits. I also agree that revising views about uniformities entails costs, depending on those views' own track record. I cannot speak to the distinction between theories as axiomatized deductive systems and as classes of models, not having read van Fraassen. Giere I have read, and I remain unclear as to how far these forms of metatalk are intertranslatable. This is partly because I count schematic diagrams and Tinkertoy models as embedding text that interprets a formalism. "Theory-rich domains" do impose tight constraints on defensive moves, and I am coming to believe the constraints on admissible *main* theories are tighter, even in the less developed sciences, than the logician's truism about "an infinite set of alternative theories" is usually taken to imply for scientists (Boyd, 1973; Meehl, 1990). (Exactly what *is* the logician's theorem for that truism, by the way? How come it does not have a name, like Gödel's, Church's, Loewenheim–Skolem, etc.? I'm wary of it. Does it hold for mathematically stated laws, or is it a trivial—and scientifically uninteresting—point about the propositional calculus? That's the only form I have seen it in: "If $r$ is a fact, we can always derive it from conjunction $p \,.\, q \,.\, r$, whatever $p$ and $q$ say.") The only place Kitcher misreads me is in interpreting "crud factor" as genuinely chance coincidences. What I, following Lykken, mean by crud factor is replicable correlations, reflecting underlying causal regularities, which in social science result in everything being correlated with everything, and hence $H_0$-refutation being usually unilluminating.

**Kukla**

1 appreciate Kukla's rendering of my argument in explicitly Bayesian terms, which should make it more acceptable to convinced Bayesians. However, very many scientists (and metatheorists!) are not Bayesians, so I preferred to keep my formulation more general. As I said in the article, non-Bayesians (e.g., Popper, Lakatos, Fisher) attach great weight to risky tests, as do working scientists who ignore metatheory and have not thought about Bayes's theorem since they took college algebra. Although Salmon thinks Bayesian, I am not persuaded one must rely on the old theorem to hold that a strong factual track record is best achieved by predicting damn strange coincidences. As I see it, the biggest single problem for the Bayesian view of theory appraisal is the allegedly infinite set of alternatives whose probabilities are summed in the second denominator term, as Kukla says. (This metatheoretical application to substantive theories does not prejudge the Bayesian position as to inferential statistics.) The nagging question about infinitely many theoretical competitors, although it surfaces brutally against Bayesians, is present for other metatheories also. It is one reason why Popper's anti-inductivism and refusal to equate corroboration with probability are attractive. Suppose that, somehow, the set of alternative theories can be treated as finite (e.g., all "otherwise admissible" theories that scientists in a domain will concoct before the sun burns out) or that, for theories using functional equations, the set is conventionally limited (Meehl, 1990). Then my selective attack on $H_0$-refutation in social science still stands, due to the weak general constraints and the large (although finite) number of plausible competitors capable of deriving a nonzero difference.

**Maxwell and Howard**

Of course I agree with Maxwell and Howard that there is an important place for inferential statistics and point-estimation techniques in psychological research. I did not intend index $C_i$ to exclude point estimation, which is highly desirable when available, as it makes the intolerance component $\simeq 1$ in the index formula. As to defective design of studies being the "main culprit," I cannot separate reliance on $H_0$-refutation from study design, because I hold that the contemplated inference from $H^*$ (mere nonnull trend) to *"T,* with good support" is, in social science, a basic mistake. My epidemiological example is weakened by realizing that a strict, "on-the-nose" result will be unlikely, but I used it because the numbers, being area rates (rather than individuals' scores), should have smaller errors; and because in that example there is no "population" of regions subject to sampling error, we have exhausted the supply. Admittedly, if the interval allowed by one's path analysis is increased to cover "near misses," its ratio to the Spielraum declines, so $C_i$ is reduced. This is not a defect, as I see it, because whenever the fact domain is numerically slippery and the theory tolerant, "successful" prediction proves less. There is just no way for us to have our cake and eat it too in these matters. Part of my complaint against conventional $H_0$-refutation is that the hypnotic fascination of "$p < .01$" diverts us from facing the hard, unavoidable trade-off. One reason (not the main one) why scientists seek qualitative diversity of experiments testing a theory is the hope that sufficient diversity will usually mean quasi-independence at the fact level, whereby the cumulative probabilities will approximate the multiplication theorem, net joint $p = p_1 \cdot p_2 \cdot p_3 \ldots p_k$. of $k$ experiments falling exponentially with $k$ even if the component $p$s must be allowed to be larger than we would like due to (a) $T$'s intrinsic tolerance and (b) allowance for statistical error (Meehl, 1990). I find myself puzzled as to just what the Maxwell–Howard "self-determined" experiment proves (it surely proves *something*), so I refrain from comment, except that I of course like it as a case of point prediction. When LISREL makes strong point (or narrow range) forecasts, it is fine with me. But my impression—shared by knowledgeable Minnesota colleagues —is that it is more commonly used as a kind of "creeping inductivism," adjusting the path diagram to progressively better fits, and of this I am suspicious. On "cursing the dark," my text contains no imprecations, but tries to say loud and clear (because I find most people won't listen) that we are in semidarkness. (I try to light a candle with $C_i$, but most of the commentators snuff it out without giving it an empirical chance to illuminate!)

### McMullin

McMullin emphasizes the other properties of good theories, and I had no intention to downplay them. Perhaps my effort at numerifying only one of them (factual fit)—and not all aspects of that one (e.g., qualitative diversity)—conveyed a wrong impression. My expectation is that all of them will someday be numerified (see following discussion), but I still insist that factual fit is ultimately decisive. Whether an index such as $C_i$ predicts the long run from the short run is an empirical question, with armchair plausibility considerations (based on the verisimilitude concept) available meanwhile. Like other theories, an empirical metatheory contains intra-theoretical derivations that make it appear more (or less, for my critics) worth investigating. I dare say that if the two kinds of factual fit $C_i$ aims to capture (point predictions and function forms) cannot be profitably numerified, the other good properties listed by Laudan, Kuhn, Kordig, and even some of the positivists will not be so either. That we currently need more detailed case histories of psychological theories I strongly agree. Whether statistical study of $C_i$'s performance must await cumulation of many such case studies I do not see as obvious, however, for reasons given in my general "statisticizing" discussion later. I conceive the actuarial/case-study division as mutually (a) stimulative, (b) cognitively suggestive, (c) confirmatory, and (d) explanatory, a view stemming from my work on the corresponding division in psychopathology. I realize that I cannot expect scholars who have not been immersed in that research to take the same view.

### Rorer

My former student Rorer provides a succinct, accurate formulation of my position; but he rejects verisimilitude, partly because we cannot "know for sure" that our theories have verisimilitude. I never said, or implied, that we could come by such certainty. But such metacertainty is not required to use the concept, just as certainty of truth is not required to legitimate *True* as a metalinguistic predicate. As Carnap pointed out against Kaufman, who made an argument similar to Rorer's for dropping 'True' from philosophy of science, if the overarching rule is to forbid terms whose predication lacks absolute certainty casewise, by Kaufman's own premises all the object-language predicates will have to be liquidated as well! We do not demand that knowledge means "know *for sure* that we know" before allowing the brave attainment word 'know' into our language (cf. discussion of the K–K postulate in Suppe, 1977, pp. 717-727). Objections to explications of verisimilitude should be based on defects of the metric, or of the logical grounds for thinking it will be correlated ($r < 1.00$, of course) with factual fit (however measured), rather than on the qualitative truth that our knowledge of the external world is not apodictic. That the semantic conception of truth avoided epistemic absolutism enabled Popper to become a scientific realist who accepts truth as a regulative ideal without being a justificationist or incorrigibilist in epistemology.

### Serlin and Lapsley

Serlin and Lapsley say "Meehl invokes the spirit of Lakatos only to deal with the problem of theory appraisal." Not so, unless we consider the strategy of Lakatosian defense to be part of appraisal, which I do not. Favorable appraisal renders Lakatosian defense rational. I do not deal with his complex doctrine of growth, which I only partly understand, and I am unsure how much I agree with it. One can be a Lakatosian about defense and its warrant without buying the whole business, some of which I fear is too much infused with Imre's (rejected) Leninism. I do not accept his dictum that the "boldness" of a theory can only be decided against the background of its available rivals. At least the boldness of a theory's *predictions* can be assessed with reference to the Spielraum. Mendel required no competing theory of heredity to see that successful prediction of backcross phenoltypic proportions was powerful evidence. Lakatos's amendments aside, I have never accepted the original Popper doctrine that antecedently improbable theories are to be preferred. Here, at least, I have always been a Bayesian. The big puzzle here lies in the difference between the *theory's* prior probability (which, like the Bayesians and the nonphilosophical working scientist, I prefer to be high) and the prior (absent theory) *predictions,* which I prefer to be low. I believe the logician's ready identification of content with consequence class is what causes the trouble. Someone more competent than I will have to fix that up. But one reason why I prefer the theory-properties list in my Figure 1 (target article) as an approach to comparing two theories' contents is that it avoids the consequence-class business, which is what killed Popper's attempt to define verisimilitude. I am more concerned with Lakatos's acceptability$_3$, as they say. As to the unreliability Lakatos adduces, *of course* "one can easily conceive of conditions which would make the estimate of verisimilitude by corroboration false." The relation, if such exists, is stochastic only (Rorer also seems to think I consider it one-to-one, a thesis that would be a form of epistemic chutzpah, if not madness). We know that a deductive or nomological relation would have to be wrong, as we know the "inductive syllogism" (Mill) *must* be wrong, because even induction by simple enumeration between observational predicates has often been in error (e.g., the platypus). That an index like $C_i$ can at best be a fallible indicator of verisimilitude (or, for an instrumentalist, of future predictive success) I took for granted, something everyone knows. I am horrified that my failure to mention this truism can suggest I thought $C_i$, or any other fact-fitting index, could have perfect validity. But in Meehl (1990), I show for some simple cases that the long-run rank-correlation between a crude measure of verisimilitude and a cruder measure of factual fit will be remarkably high. As for new, bold theories with no money in the bank yet, I give no rules, because (a) I don't know how and (b) I don't see why we need any. We do not have to "accept" or "reject" a new theory before it has been put to *any* predictive tests, do we? Of course the *non*factual properties mentioned earlier may properly play a role, sometimes determinative. A theory of mitosis would have been rejected out of hand if it postulated fine silver wires as components of the spindle. Nor am I Lakatosian as to excess content, because theories have been profitably researched despite their not handling some of the "old facts" that a predecessor could handle. I believe this strategic question will turn out to be much more complicated than Popper or Lakatos (or anyone else) has explained.

## The Corroboration Index, Verisimilitude, and Statisticizing Metatheory

Although all commentators agree with my overall position in its critical aspects, almost all oppose the corroboration index idea, and none of them waxes enthusiastic about it. Defects in $C_i$'s standardization (e.g., possible negative values as shown by Kitcher) can be repaired by suitable convention, or left as is. Some of the objections were anticipated and, I believe, answered in my article. To some I have no satisfactory reply, especially under a time deadline and space limitation. I think it best to address the core problem, pervading the complaints and clearly *not* springing from the critics' numerical-statistical worries about $p$ values, tolerances, metric, standardization, sampling, Spielraum specification, and so forth. If the basic idea of $C_i$ is sound, these technicalities are up for formal *and empirical* study. If the whole notion is inherently bad, we need not argue about the statistical details. (For example, Dar—whose previous excellent article on these matters was sympathetic to my critical side—while raising some important questions about the numerification proposed, labels the corroboration index "meaningless," a metalanguage epithet I thought had gone out with the death of positivism. Has the ghost of 1930 Vienna reappeared in Tel Aviv?) The easiest exposition is by succinct summary statements, not argued or referenced, either because the case was made in my article, or because I am making it with more space (and thought!) in forthcoming works (Meehl, 1990, [1992]). I number the theses for convenient reference.

1. All empirical sciences that command our assent and esteem tend to become more quantitative, both at the observational and theoretical levels, as they advance. Are there good reasons for expecting metatheory to take a different developmental course? There may be, but I have not heard of them.

2. Scientific theories are appraised by several attributes, lists having been offered by Laudan, Kordig, Kuhn, Salmon, and even the logical positivists. Sometimes these criteria pull oppositely. Disagreement persists as to their relative importance. "Factual fit," however, is ultimately decisive.

3. Scientists are impressed with factual fit when the theory's predictions are (a) narrow ("risky") and (b) accurate ("hit" or "near miss"). So it is reasonable to start with a risky-accurate composite in concocting a factual-fit index. This my $C_i$ aims to provide.

4. Scientists and metatheorists regularly employ terms of quantity in nonstatistical metadiscourse (e.g., "typical," "marked," "improbable," "more important than," "frequently," "close," "by and large," "extreme," "balances," "strongly," "normally"'). There is no argument or evidence in psychology to show that explicit numerification of these *intrinsically* quantitative claims tends to disadvantage.

5. A large body of empirical research (some 150 studies in human outcomes prediction alone) shows that humans are markedly inefficient at integrating data, so that even crude, non-optimizing formal indices (e.g., an unweighted linear composite of relevant variables) do as well or better than "skilled judges." I am confident that this point is insufficiently appreciated by my critics (except Rorer?), and I earnestly entreat them, and readers, to study the works of Dawes (1988), Faust (1984), Kahneman, Slovic, and Tversky (1982), Mahoney (1976), and Nisbett and Ross (1980) on this crucial premise of my argument.

6. Some theories are better than others, and every scientist proceeds on that basis. For a scientific realist, "better" means "closer to the truth." Despite Popper's earlier failure at an explication, people are working on it (Goldstick & O'Neill, 1988; Meehl, 1990; Newton-Smith, 1981; Niiniluoto, 1984, 1987; Oddie, 1986; Tichý, 1978). But I think the approach in my Figure 1 is better than the logicians'. Would Rorer, who dislikes the concept, say that if $T_1$ and $T_2$ differ only at Level IX (numerical values of function parameters), whereas $T_1$ and $T_3$ differ at all levels, starting with the kinds of entities postulated, we can attach no meaning to the metacomment "$T_2$ is more similar to $T_1$ than $T_3$ is to $T_1$" ? I cannot conceive he would say that. As to the metatheoretical derivation of verisimilitude's stochastic linkage to factual fit, an adequate development is impossible in the space available, so I must refer the reader to Meehl (1990); but herewith an example. In the MacCorquodale–Meehl formulation of expectancy theory (MacCorquodale & Meehl, 1953, 1954; Meehl & MacCorquodale, 1951), the conjectured "mnemonization postulate" makes an expectancy ($S_1R_1S_2$) grow as a monotone increasing decelerated function of the number of close-contingency ($S \rightarrow R_1 \rightarrow S_2$) sequences run off by the rat. Suppose there are no such entities in the rat's brain as Tolmanian expectancies (as Watson, Hunter, Guthrie, Hull, or Skinner would say). The mnemonization postulate is in the "hard core" of cognitive theory, pervading the nomological network, and occurring essentially in almost all derivation chains to theoretical well-formed formulas (coordinated "operationally" to observational well-formed formulas). It is a Level I error in my theory property list (Figure 1), and almost all observational consequences obtainable by conjoining it with various subsets of the other postulates will be found false in the lab. Suppose it were qualitatively correct but the function, while monotone increaseing, is linear rather than decelerated, an error at Level III. Many experiments of the conventional kind, testing for "an effect" (even Fisherian interactions) but not attempting to fit a function

(e.g., log $n$ or $1 - e^{-kn}$), will pan out. But those experiments that do fit a function form will not fit the deceleration conjecture. Now imagine that all but one of our postulates are literally correct, the functions *with parameters* being filled in theoretically; so everything agrees with Omniscient Jones's true theory except, say, a small parametric error in Postulate 7, *induced elicitor-cathexis:*

> The acquisition of valence by an expectandum $S_2$ belonging to an existing expectancy ($S_1R_1S_2$) induces a cathexis in the elicitor $S_1$, the strength of the induced cathexis being a decelerated increasing function of the strength of the expectancy and the absolute valence of $S_2$. (MacCorquodale & Meehl, 1954, p. 244)

Only a few experimental designs (aimed at detecting elicitor cathexis) will come out wrong, and these only by a small quantitative deviation, because the postulate is correct up through signs of derivatives, function forms, and transsituationality of parameters, erring only at Levels VIII and IX. Examples like this suffice to show ("logically") that verisimilitude and a factual-fit statistic—however crude—will be correlated. Verisimilitude is an absolutely necessary metaconcept for *both* the scientist and the metatheorist, and we just have to keep working on its explication. I am puzzled that a bunch of postpositivists are so intolerant of making do with open concepts in a research program aimed to tighten them. As Campbell says, one of the liberating results of Cronbach and Meehl (1955) and MacCorquodale and Meehl (1948) was their *open-concept permissiveness.* I cannot refrain from pointing out that some of the most fundamental terms in science are still inadequately explicated, whether by scientists or philosophers. Many writers have noted that the most basic and pervasive notions are often hardest to define rigorously. One thinks of such concepts as observable, probability, randomness, causal nexus, dispositions, counterfactuals, partial interpretation, reduction, confirmation, implicit definition, and analyticity.

7. If a theoretical entity or property $\theta$ is inaccessible directly, but alleged to be accessible indirectly via an accessible $x$, this indirect accessibility relies on a lawlike (nomological or stochastic) relation between $\theta$ and $x$. But how can such a relation be verified, since the relata are not independently accessible? It seems like some circularity must be involved. Well, yes and no. As Feyerabend once said to me—one of his provocative sallies containing a deep truth—"There's nothing wrong about arguing in a circle if it's a big enough circle." As is well known, this is the rock on which foundationalist phenomenalism founders in general epistemology. With only the single ($\theta \rightarrow x$) linkage, it can't be done. What happens, of course, is that $\theta_1$ is also linked to $\theta_2$, which is in turn linked to accessible $y$, and so on within a *law network,* in which Popper's "basic statements" (privileged but corrigible) about $x$ and y find their place. The accessible relations among ($x, y, z, \ldots$) corroborate the conjectured network that includes the θs. *I hold that the relation between verisimilitude and the familiar set of good properties of theories is closely analogous to that of first-level theories to their corroborating facts.*

8. What this means for our problem I described briefly in the article. One constructs various quantitative indices of "good" theory properties. Their desired linkage to verisimilitude is evidenced in three interlocking ways: (a) theoretical derivation, at least for simple idealized cases, as in Meehl (1990); (b) discriminant analysis between high-confidence true and false theories (it's harmless if a small fraction of theories classified true are later rejected—the relation is stochastic, and the statistical situation is similar to that of psychometric item-analysis against a fallible diagnostic "criterion"; cf. Cronbach & Meehl, 1955; Golden & Meehl, 1978; Meehl & Golden, 1982, on the bootstraps effect); and (c) factor analysis of the indices' correlation matrix, followed by matching the factor-loading profile with the discriminant weights of (b). Why did none of the commentators discuss this powerful construct-validating approach? We do not demand a deductive demonstration that a composite index *must* correlate strongly with verisimilitude, because we reject the "K–K principle" that you cannot have knowledge without knowing with certainty that you have it (Hintikka, cited by Suppe, 1977, pp. 716-728). We give plausibility arguments that it will, but the test is empirical. There is a deep sense in which correspondence theorists rely on coherence; that applies here as well. If the *set* of indices "works" empirically in this convergent stochastic sense, and a fact-fit index like $C_i$ does its job well, the objections of my critics will have been refuted, *modus tollens.* If $C_i$ does poorly, their pessimistic conjectures are corroborated.

When 14 able minds are so unenthusiastic about my index proposals, why don't I capitulate? Several reasons. First, as a neo-Popperian, I do not think it a sin or disgrace to be wrong in a bold conjecture. Second, in my work as a psychologist, I have a history of being in a small minority but turning out to be correct years later (e.g., superiority of structured tests over projectives, schizophrenia as a neurological disorder, actuarial vs. clinical prediction, inefficacy of psychotherapy for criminals, merits of Albert Ellis's rational emotive therapy, cognitive [expectancy] theory of animal learning, genes and alcoholism, construct validity in psychometrics, importance of heredity in intelligence and personality, the value of taxonic nosology in mental disorders). So being *vox clamantis in deserto* doesn't bother me. Third, I suspect few of the critics have been steeped in the judgment literature, as I have. One needs that for perspective. Fourth, for more than a third of a century, I have observed the determined refusal of psychologists to admit the actuarial thesis in the face of massive, diverse, and consistent research evidence (Dawes, Faust, & Meehl, 1989; Meehl, 1986). It is apparently an extremely hard notion for humans to assimilate. Fifth, we know from history of science that radically novel ideas regularly meet with resistance, and statisticizing metatheory is certainly a new—and radical—idea.

As to $C_i$'s quantitative imperfections, I trust some are correctible (e.g., decelerate the metric? adjust standardizing constants?), whereas others we would learn to live with, as we do with IQ, windchill factor, consumer price index, uniform crime reports, Hollingshead socioeconomic status, and World Health Organization indices of quality of life. In employing any useful numerification of an open concept in the social sciences, one is properly alert to the caveats, but not frightened into cognitive paralysis by them. (When serving on a National Research Council committee on criminal deterrence, I was told by a distinguished economist that we should not even discuss Sellin's severity index of criminality, absent rigorous formulation and proof that the seriousness of different crimes can be located on an interpersonal cardinal utility metric. So the taxpayer's view that a rape and two armed robberies makes an offender more scary than three shopliftings is meaningless. Is such mathematical purism reasonable? I think not.) As to the danger of scientists' overemphasizing $C_i$ to the neglect of other important aspects of theory, I must first invoke the medieval moralists' *abusus non tollit usum* (the abuse does not destroy the use). Secondly, I conjecture that other theory properties will also be amenable to numerification, so the seductiveness of "having a number to look at" will be equalized. Thirdly, I confidently predict—from 36 years experience of the clinical-statistical controversy in my own science—that most persons are more likely to be skeptical, or even hostile, to numerification than attracted by it—witness my critics!

The same rejoinders are appropriate with respect to verisimilitude, both its explication and its hoped-for correlation with factual track record, whether indexed by $C_i$ or otherwise. We must keep working at it, and my article was intended simply as a contribution to that collective effort. That there will be *some* correlation between fact-fitting track record and verisimilitude is quite easy to show, even with crude measures of both concepts (Meehl, 1990).

But I discern, in all but a couple of my critics, resistances more fundamental and pervasive than these concessions, bufferings, and rejoinders can meet. I gather that almost all of them reject the idea of *any* such statisticization of theory performance, or that it could ever be shown to correlate with verisimilitude, or both. I am not sure just how to deal with this sort of flat rejection, which seems to be saying, "We should not even *try* to do this, because we know it *can't* succeed, so why waste time, brains, and energy fooling around with it?" Because my rationale, as a neo-Popperian, for offering conjectures is that we have a problem, I take it that my critics either (a) deny we have a problem or (b) know that my conjecture cannot possibly be adequate to solve it. I confess I do not understand how they can be confident of either (a) or (b).

It may be debatable whether scientists themselves have a problem in assessing theories, but I have advanced evidence and arguments to show that they do. It puzzles me that my critics did not address themselves to the sizable body of research on human malcognition. I am perhaps hyperaware here, because my expertise on the clinician's errors leads me to be skeptical about scientists, seeing that the psychology, sociology, statistics, *and epistemology* of the diagnostic and prognostic process (whether in organic medicine, psychopathology, criminology, personnel selection, sports forecasting, business decisions, personality attribution, or whatever) is similar in almost all respects to that of assessing a scientific theory from a complicated and often inconsistent mass of evidence. As I argued in the article (to no critic's denial, I am pleased to note), that science is—usually and in the long run—a more successful cognitive enterprise than ethics, aesthetics, metaphysics, theology, literary criticism, or theoretical historiography tells us *nothing* about how close it is to cognitive optimality. But even if it were held that the scientist, proceeding informally, cognizes with near maximum efficiency, surely no one will urge that philosophy of science is proceeding smoothly and rapidly to consensus! Almost every thesis of postpositivist metatheory is in dispute, even the definition of its task and methods. When we have PhDs with high IQs and knowledge of the sciences ranging in viewpoint from Paul K. Feyerabend to Carl R. Kordig, things are in pretty much of a mess; and the role of factual adequacy is certainly not among the "nonmessy" areas, if there are any such.

Assuming *arguendo* that metatheory presents difficult problems, I conclude that my critics think we can say *today* that an index such as $C_i$ will fail to help, that verisimilitude is an inadmissible concept, and that the relation between $C_i$ and verisimilitude is absent (or, at least, unprovable). That is, they reject a conjectured problem-solver on an armchaired certainty of failure. I take my former student Rorer (who is much disposed in my favor on most matters) as an example. He writes, concerning the concept of verisimilitude and its postulated correlation with evidentiary support, "The reason no philosopher of science has done that, I submit, is that it can't be done." *How does Rorer know this?* How can he, or anybody, come by such high-certainty forecasting of future developments in metatheory? This seems a strange a priorism to find in a postpositivist thinker, does it not? In the old days of my positivist youth, it might have had some warrant, when the linguistic turn was in the ascendant. In his *Logical Syntax of Language,* Carnap (1934/1937) set philosophy of science = logic of science = logical syntax of scientific language = combinatorics of certain geometrical shapes (Neurath's "mounds of ink"), and, on that hyperlinguistic view, one may suppose most questions, no new empirical facts being needed or relevant, should be readily soluble. But not all, even on that discarded theory of metatheory. Purely formal sciences have problems that remain unsolved for long periods of time. Today mathematicians do not know the truth about Fermat's last theorem, Goldbach's conjecture, the Riemann zeta hypothesis, or Cantor's continuum conjecture. On this last, almost a half century elapsed before Gödel (in 1938) proved it was consistent with set theory, and then more than a quarter century before Paul Cohen (in 1963) showed its contradictory was also (Cohen & Hersh, 1967; cf., more technically, Cohen, 1966). The time lapse since Popper introduced verisimilitude is small by comparison. And what is true even for purely formal sciences of course holds a fortiori for empirical disciplines. I am not troubled in the least by formal metaproofs against the comparability of two false theories, because—as I argue in the article—I reject the logician's approach to it (in terms of consequence class, etc.). I note that my very different approach (see Figure 1 in target article) in terms of increased specification of a theory's quantification properties was not examined by the critics, which I find puzzling in those who find the very concept of verisimilitude objectionable.

Assume *arguendo* that I am correct in my belief that the critics err in armchair rejection of a conjecture aimed to approach solution of a real problem. How came these able, learned, and kindly disposed men to be thus mistaken? I do not argue ad hominem in asking this, for conjecturing as to a possible cognitive source of error, once admitted, is not like attributing unproved error by imputation of motive or by production of social embarrassment. It would be interesting to inquire how the meaning of 'argumentum ad hominem' will have to be restricted in a metatheory that admits psychosocial facts—more, in many postpositivist thinkers, assigning them a principal role! (For an illuminating and unsettling discussion of the poorly defined ad hominem fallacy from Aristotle to the present, see Hamblin, 1970).

So I offer an interpretation of why my critics went awry, in an irenic and clarifying spirit. I conjecture that both (a) their lack of appreciation for the problem of informal, non-statistical theory appraisal and (b) their armchair rejection of my proposed partial solution, stem from the same underlying cognitive defect: *They have not fully assimilated the postpositivist view of metatheory as the empirical theory of theories.* Despite their being more happily "empirical" about metatheory than I, an old ex-positivist, they have not perceived the new metatheory's implications for method as fully as I (reluctantly) have done. This is a strong (and psychoclinical?) kind of thesis, but let me try to defend it. Of course, I do not here suggest anything hypocritical or even disingenuous; merely an understandable failure to perceive the ramifications of the new doctrine.

If metatheory is an empirical science, it will presumably live the life that other empirical sciences live, although perhaps differing (in degree but not in kind) by virtue of its generality. This means we can confidently expect it to undergo amendment, expansion, problem shifts, surprises, disappointments, doldrums, conjectures, and refutations, and a variable interplay between formal and factual considerations. It will permit idealizations and approximations, some rougher than others. It will tolerate open concepts, while endeavoring to tighten them, chiefly by statistical methods but also by semantic revisions. Following Carnap, it will have a principle of tolerance, and will offer "explications" of preanalytic intuitions more often than rigorous "definitions" of its concepts. That it is avowedly empirical, based on case studies (and, if my view prevails, multiple statistical indicators) of the history of science, does not imply that it is devoid of formal arguments, any more than physics, economics, or cognitive psychology eschew mathematics and logic because they are empirical. What does such a picture of metatheory mean? First, it means that one cannot confidently foresee the course of development. (Popper amusingly pointed out that if the determinist-historicist could predict the course of physics, then he could "do physics" without being a physicist, which is absurd.) So the properties of a fact-adequacy index like $C_i$ are investigated by a combination of formal and empirical approaches. New efforts to explicate 'verisimilitude' will be similarly subject to both kinds of scrutiny. On such a view, Rorer cannot conceivably know in 1990 whether, or when, some logician will explicate 'verisimilitude' in a satisfactory way; nor can I. One simply cannot accept the postpositivist view of metatheory as an empirical discipline and then proceed to dogmatize about its future course. As Feyerabend (1970) pointed out, even the basic principle that we can at least forbid out-and-out logical contradictions *within* a theory is not always adhered to, as it should not be once we substitute truth-likeness for truth. Employing contradictory concepts at certain stages of a science has sometimes been helpful (e.g., the Bohr atom), and for the same reasons that admittedly false idealizations have been transitorily indispensable (e.g., gas molecules as perfectly elastic point-masses in deriving the gas law from kinetic theory). It is only on a purely linguistic view that one can "settle" a metatheoretical question by sheer taking thought, without the trial-and-error of an empirical science.

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?"

The target article was cynical about most psychological theories and challenged the conventional method of appraising them, but went on to suggest an alternative approach. Because the commentators generally agree with the former but reject the latter, the net result may seem pessimistic. About my own main field (clinical psychology), I must admit considerable "cognitive disappointment" (Meehl, 1989). Yet I persist in long-term optimism about even this "soft" area. It has five noble intellectual traditions that I am sure will survive and improve: (a) psychodynamics, (b) descriptive psychopathology and nosology, (c) applied learning theory, (d) behavior genetics, and (e) psychometrics (Meehl, 1987). Sigmund Freud, a great contributor to the first two, was crystal clear (and optimistic) about open concepts and their gradual explication by the research process:

> We have often heard it maintained that sciences should be built up on clear and sharply defined basic concepts. In actual fact no science, not even the most exact, begins with such definitions. The true beginning of scientific activity consists rather in describing phenomena and then in proceeding to group, classify and correlate them. Even at the stage of description it is not possible to avoid applying certain abstract ideas to the material in hand, ideas derived from somewhere or other but certainly not from the new observations alone. Such ideas—which will later become the basic concepts of the science—are still more indispensable as the material is further worked over. They must at first necessarily possess some degree of indefiniteness; there can be no question of any clear delimitation of their content. (1915/1957, p. 117)

A very different sort of powerful intellect was Edward Lee Thorndike, a fertile thinker and investigator in the other three traditions. Having the courage of his quantifying convictions, he attached useful numbers to such unlikely things as handwriting quality, personal values, and the goodness of cities. I cannot trace the reference, but I memorized this passage as a student; he wrote:

> Our ideals may be as lofty and as subtle as you please. But if they are real ideals, they are ideals for achieving something; and if anything real is ever achieved, it can be measured. Not perhaps now, and not perhaps 50 years from now. But if a thing exists, it exists in some amount; and if it exists in some amount, it can be measured.

## Note

Paul E. Meehl, Department of Psychology, N218 Elliott Hall, University of Minnesota, 75 East River Road, Minneapolis, MN 55455.

## References

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

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

Carnap, R. (1936-1937). Testability and meaning. *Philosophy of Science, 3,* 420-471; *4,* 2-40. (Reprinted with corrigenda and additional bibliography, New Haven, CT: Yale University Graduate Philosophy Club, 1950; and in H. Feigl & M. Broadbeck [Eds.], *Readings in the philosophy of science* [pp. 47-92] New York: Appleton-Century-Crofts, 1953.)

Carnap, R. (1937). *The logical syntax of language* (A. Smeaton, Trans.). New York: Harcourt, Brace. (Original work published 1934)

Carnap, R. (1950). Empiricism, semantics, and ontology. *Revue Internationale de Philosophie, 4,* 20-40.

Carnap, R. (1956). The methodological character of theoretical concepts. In H. Feigl & M. Scriven (Eds.), *Minnesota studies in the philosophy of science, Vol. I: The foundations of science and the concepts of psychology and psychoanalysis* (pp. 38-76). Minneapolis: University of Minnesota Press.

Cohen, P. J. (1966). *Set theory and the continuum hypothesis.* New York: Benjamin.

Cohen, P. J., & Hersh, R. (1967). Non-Cantorian set theory. *Scientific American, 217,* 104-116.

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* [pp. 3-31]. Minneapolis: University of Minnesota Press, 1973)

Dawes, R. M. (1988). *Rational choice in an uncertain world.* Chicago: Harcourt Brace Jovanovich.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science. 243,* 1668-1674.

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

Feigl, H. (1950a). Existential hypotheses: Realistic versus phenomenalistic interpretations. *Philosophy of Science, 17,* 35-62.

Feigl, H. (1950b). The mind-body problem in the development of logical empiricism. *Revue Internationale de Philosophie, 4,* 64-83.

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* (pp. 17-130). Minneapolis: University of Minnesota Press.

Freud, S. (1957). Instincts and their vicissitudes. In J. Strachey (Ed. & Trans.), *The standard edition of the complete psychological works of Sigmund Freud* (Vol. 14, pp. 117-140). London: Hogarth. (Original work published 1915)

Golden, R., & Meehl, P. E. (1978). Testing a single dominant gene theory without an accepted criterion variable. *Annals of Human Genetics London, 41,* 507-514

Goldstick, D., & O'Neill, B. (1988). "Truer." *Philosophy of Science,* 55, 583-597.

Hamblin, C. L. (1970). *Fallacies.* London: Methuen.

Hempel, C. G. (1950). Problems and changes in the empiricist criterion of meaning. *Revue Internationale de Philosophie, 4,* 41-63.

Kahneman, D.,Slovic, P.,& Tversky, A. (Eds.). (1982). *Judgments under uncertainty: Heuristics and biases.* Cambridge, England: Cambridge University Press.

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin, 70,* 151-159. (Reprinted in D. E. Morrison & R. E. Henkel [Eds.], *The significance test controversy* [pp. 267-279]. Chicago: Aldine, 1970)

MacCorquodale, K., & Meehl, P. E. (1948). On a distinction between hypothetical constructs and intervening variables. *Psychological Review, 55,* 95-107.

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* (pp. 177- 266). New York: Appleton-Century-Crofts.

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

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence.* Minneapolis: University of Minnesota Press.

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* (pp. 349-411). Minneapolis: University 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* (pp. 7-23). Norman, OK: Transcript Press.

Meehl, P. E. (1989). Autobiography. In G. Lindzey (Ed.), *History of psychology in autobiography* (Vol. 8, pp. 337-389). Stanford, CA: Stanford University Press.

Meehl, P. E. (1990). *Corroboration and verisimilitude: Against Lakalos' "sheer leap of faith"* (Working paper). Minneapolis: Minnesota Center for Philosophy of Science.

Meehl, P. E. (1992). Cliometric metatheory: The actuarial approach to empirical history-based philosophy of science. *Psychological Reports, 71*, 339-467. [reference updated]

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

Meehl, P. E., & MacCorquodale, K. (1951). Some methodological comments concerning expectancy theory. *Psychological Review, 58,* 230-233.

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

Niiniluoto, I. (1984). *Is science progressive?* Boston: Reidel.

Niiniluoto. I. (1987). *Truthlikeness.* Boston: Reidel.

Nisbett, R. E., & Ross, L. (1980). *Human inference: Strategies and shortcomings of human judgment.* Englewood Cliffs, NJ: Prentice-Hall.

Oddie, G. (1986). *Likeness to truth.* Boston: Reidel.

Pap, A. (1953). Reduction-sentences and open concepts. *Methodos, 5,* 3-30.

Suppe, F. (Ed.). (1977). *The structure of scientific theories* (2nd ed.). Urbana: University of Illinois Press.

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

Waismann, F. (1945). Verifiability. *Proceedings of the Aristotelian Society, 19,* 119-150.


pdf by ljy May 2003