*PAUL E. MEEHL**

THE MIRACLE ARGUMENT FOR REALISM: AN IMPORTANT LESSON TO BE LEARNED BY GENERALIZING FROM CARRIER'S COUNTER-EXAMPLES

SUPPOSE the Miracle Argument for scientific realism is formulated: If the entities postulated by a theory *T* do not exist, if there are no such entities having the properties the theory attributes to them, then *T* cannot successfully predict new observational facts. If "facts" is taken to mean "at least one fact", it follows rigorously from that formulation that if *T* manages to predict a single novel fact, *T* must be true. Carrier[1] shows, by two clear counter-examples from the history of chemistry and physics, that this metatheoretical principle is unsound. The theories of phlogiston and caloric, known today to be false, each made a novel quantitative prediction. (The former made two successful predictions, the latter one or two, depending on how formulated).

The present paper is not a reply to Carrier, but a development of some consequences. However, I cannot avoid the question whether those philosophers who subscribe to the Miracle Argument hold it in the strong form stated above, which is the form Carrier refutes. If they do, they should not, since the strong form seems to render inductive inference infallible under suitable circumstances, a single observational fact admitted into the corpus being dispositive. One of the few metatheoretical principles universally accepted as "settled" is that empirical generalizations from factual particulars, being ampliative, are at best probable. If this holds even for "first level" inductions consisting of observation terms, it would be strange to find it not holding for generalizations about theoretical entities known to us only by indirection of a complicated epistemic kind. I know from conversation and correspondence, as well as from their writings, that neither Giere, Popper, nor Salmon — all scientific realists relying on the Miracle Argument — holds it in the strong form, making a predictive "success" dispositive as to truth. One may doubt whether any philosopher (or reflective scientist) has held its strong form.

*Member of the Minnesota Center for Philosophy of Science. Address: Department of Psychology, N218 Elliott Hall, 75 East River Road, Minneapolis, MN 55455, U.S.A.

Received 29 *August* 1991; *in revised form* 14 *November* 1991.

[1]M. Carrier, 'What is Wrong With the Miracle Argument?' *Studies in History and Philosophy of Science* **22** (1991), 23–36.

Stud. Hist. Phil. Sci., Vol. 23, No. 2, pp. 267-282, 1992

Carrier does not restrict the Miracle Argument to novel predictions, as Whewell's "consilience of inductions" covers cases where a previously known but unexplained fact is seen to be handled by a theory not concocted for that purpose.[2] And he is of course aware that scientific realists do not usually jump from a *single* novel fact prediction to the reality of the theoretical entities, rather they make the weaker and safer claim that empirically successful theories are probably approximately true.[3] It is unfortunate and confusing that some scientific realists employ the term miracle to make their realist point forcefully. I dare say their world view is such that they disbelieve in miracles, that is, *they assume miracles do not in fact ever happen*. (Perhaps they would even be willing to apply the causal modality *impossible* to alleged miracles?) This very strong "cannot happen" flavor is doubtless the reason for their choosing the term 'miracle' in a nontheological context. But the connotation comes dangerously close to the "once-is-enough" strong thesis that Carrier does not attribute to them. And it seems odd that if one novel prediction cannot clinch it, nor 2, nor 3, there is some number *k* of empirical successes that would be "miraculous" (i.e., could not happen) were the theory objectively false.

It is worth mention that although no philosopher would accept the once-is-enough form, and probably no scientist would do so if challenged in a metatheoretical discussion, one does read and hear scientists expressing their confidence *as if* they subscribed to it. I have rather frequently found scientists claiming that a certain prediction (typically a "risky" one) "settles it", "clinches it", "cannot be explained otherwise", "is a clear proof", "is definitive", "provides unassailable support", "is overwhelming evidence". (Whenever I hear a scientist say "overwhelming", I become suspicious. Nobody ever needs to say the evidence against a flat earth is overwhelming.)

Let us try a less grandiose statement of the idea: If *T* successfully predicts novel facts, this track record is a good reason for thinking that the entities postulated by *T* exist and have, *more or less* (or, approximately), the properties *T* attributes to them, i.e., *T* has considerable verisimilitude. A corollary might be that the more such predictive "hits" *T* achieves, and the smaller their antecedent probabilities (absent *T*), the better; so that if *T* achieves a large number of such "damn strange coincidences"[4] it is reasonable *until further notice* to ascribe verisimilitude to *T*. Further, for a given theory and fact

[2]*Ibid*., pp. 26–27.

[3]*Ibid.*, p. 35, and Carrier, personal communication, 21 October 1991.

[4]W. C. Salmon, *Scientific Explanation and the Causal Structure of the World* (Princeton, NJ: Princeton University Press, 1984); and personal communication June 1980.

pattern ($T \rightarrow f_1, f_2, f_3, f_4$), we would assign more weight if *T* had *predicted* novel facts f_3, f_4 (*T* having been invented to *explain* f_1, f_2) than if *T* were concocted with f_1, f_2, f_3, f_4 before us. That is, the mixed argument from convergence and prediction receives more probative weight than the pure argument from convergence.[5] Philosophers disagree about this second corollary. I recall Carnap saying, "But, Meehl, how can the mere *date* of a fact affect its logical relation to a hypothesis?" On the other side, one has the impression that Giere and Popper ascribe *no* (or negligible) value to pure convergence. I suspect most metatheorists are (like me) in between, as are almost all the scientists (physical, biological, social) I have questioned. A common view is that "mere" after-the-fact explanatory convergence, with no novel predictions, deserves some weight when the facts are of a numerical kind specifying a narrow quantitative range; but that, *ceteris paribus*, the more predicted, the better, i.e., prediction outweighs convergence.[6] Doubtless the positivists and Popper were uncomfortable about this preference because of their fear of committing the sin of "psychologism". If we conceive metatheory as the empirical theory of scientific theories, hence, as a branch of social science,[7] we are less worried about psychologism. I have offered a proof, based on statistical theory plus a bare minimum of psychologism, that the working scientist's liking for prediction over convergence is a rational preference.[8]

Restricting attention to scientists and metatheorists who subscribe to the weaker form, it being the only one defensible, do we consider successful risky prediction the *sole* consideration appropriate to theory appraisal? I know of no such thinkers, and I am prepared to wager that there aren't any, as I dare say the reader would also. Such a single touchstone, a purportedly perfect litmus test, of theoretical verisimilitude doesn't fit the practice of scientists, nor does it have metatheoretical justification. We need not buy anyone's favorite list of properties to agree that scientists typically take several things into account

[5]A. Castell, *A College Logic* (New York: Macmillan, 1935).

[6]But cf. S. G. Brush, 'Prediction and Theory Evaluation: The Case of Light Bending', *Science* **246** (1989), 1124–1129.

[7]J. D. Sneed, 'Philosophical Problems in the Empirical Science of Science: A Formal Approach', *Erkenntnis* **10** (1976), 115–146; Sneed, *The Logical Structure of Mathematical Physics*, 2d edn (Boston: D. Reidel, 1979).

[8]P. E. Meehl, *Corroboration and Verisimilitude: Against Lakatos' "Sheer Leap of Faith",* (Working Paper, MCPS–90–01) (Minneapolis: University of Minnesota, Center for Philosophy of Science, 1990), see pp. 11–15.

when appraising a theory. Many "lists" of desirable features have been set out[9] — there are doubtless scores of such partly overlapping indicators of theory merit in textbooks, treatises, and articles. All I need here is that several different features are normally invoked; reader, pick your own favorites.

When one contemplates such a list, it is obvious that the members cannot be derived from one another; nor has anyone claimed to derive his list from a single metatheoretical principle or concept. They are all "nice properties to have" in some intuitive sense, but do not flow from any root property. So we should expect examination of theories in history of science to reveal instances of countervailing properties, a theory "looking good" when judged by properties P_1, P_2, …P_m but not so good in light of P_n, P_{n+1},…. This is, of course, what we find, and this mixed finding is so ubiquitous that no statistical study is needed to be confident of the meta-generalization. Not as "proof" but illustrative, I offer a striking example from psychology. Skinner's operant behavior theory is experimentally based (lever pressing in the Skinner box by rats and pigeons) and shows beautiful replicability and considerable quantitative detail *in that context*. How well it "works" in the broader domain (e.g., rats in the maze, adult human social learning, or children's language acquisition) is disputed and largely programmatic. Freud's theory impresses us with the great breadth of its explanatory power (neurotic symptoms, character traits, dreams,

[9]E.g., M. R. Cohen and E. Nagel, *An Introduction to Logic and Scientific Method* (New York: Harcourt, Brace, 1934), see pp. 207–215; I. M. Copi, *Introduction to Logic*, 2nd edn (New York: Macmillan, 1961), see pp. 426–433; F. W. Dauer, *Critical Thinking: An Introduction to Reasoning* (New York: Oxford University Press, 1989); D. Faust and P. E. Meehl, 'Using Scientific Methods to Resolve Enduring Questions Within the History and Philosophy of Science: Some Illustrations', *Behavior Therapy* **23** (1992), 195-211; H. Feigl, 'Meaning and Validity of Physical Theories', (original publication in German, 1929) translated and reprinted in R. S. Cohen (ed), *Herbert Feigl: .Inquiries and Provocations: Selected Writings 1929–1974* Boston: D. Reidel, 1981), pp. 116–144, see pp. 131–137; H. Feigl, 'Existential Hypotheses', *Philosophy of Science* **17** (1950), 35-62, see pp. 38–41 (Reprinted in R. S. Cohen (ed), *ibid*., pp. 192–223, see pp. 196–200); C. G. Hempel, *Philosophy of Natural Science* (Englewood Cliffs, NJ: Prentice-Hall, 1966), see pp. 33–46; C. R. Kordig, *The Justification of Scientific Change* (Boston: D. Reidel, 1971); Kordig, 'The Comparability of Scientific Theories', *Philosophy of Science* **38** (1971), 467–485; Kordig, 'Discovery and Justification', *Philosophy of Science* **45** (1978), 110–117; T. S. Kuhn, 'Objectivity, Value Judgment, and Theory Choice', in *The Essential Tension: Selected Studies in Scientific Tradition and Change* (Chicago: University of Chicago Press, 1977), pp. 320–339, see pp. 320ff (reprinted in B. A. Brody and R. E. Grandy (eds), *Readings in the Philosophy of Science*, 2nd edn (Englewood Cliffs, NJ: Prentice Hall, 1989), pp. 356–368, see pp. 356ff); L. Laudan, *Progress and Its Problems: Toward a Theory of Scientific Growth* (Berkeley: University of California Press, 1977); Laudan, *Science and Values* (Berkeley: University of California Press, 1984); H. Margenau, *The Nature of Physical Reality* (New York: McGraw-Hill, 1950), see pp. 81–121; W. H. Newton-Smith, *The Rationality of Science* (Boston: Routledge & Kegan Paul, 1981), see pp. 226–232; K. R. Popper, *Conjectures and Refutations* (New York: Basic Books, 1962), see pp. 231–233; K. F. Schaffner, 'Outlines of a Logic of Comparative Theory Evaluation with Special Attention to Pre- and Post-Relativistic Electrodynamics', in R. Stuewer (ed), *Minnesota Studies in the Philosophy of Science: Vol. V. Historical and Philosophical Perspectives of Science* (Minneapolis: University of Minnesota Press, 1970), pp. 311–354), see pp. 318–330; D. Shapere, 'Scientific Theories and Their Domains', in F. Suppe (ed), *The Structure of Scientific Theories*, 2nd edn (Chicago: University of Illinois Press, 1977), pp. 518–589.

jokes, slips, fairy tales, myths, literary criticism, anthropology, even primate ethology) but is non-experimental, minimally quantified, and excessively subject to *ad-hock*ery. Which theory has more going for it? Psychologists continue to disagree.

Given that the dozen or more theory properties are individually nondispositive and are (we hope) *stochastically* related to verisimilitude, hence, often in opposition to one another, how should theory appraisal proceed? We are faced with a problem of *combining factors*, which means that we are, willy-nilly, counterbalancing some factors with others to reach a net evaluation. However we manage this, we will be *assigning weights* (even if all the weights are equal) and reaching a composite "score" of judged merit. The working scientist may not (typically, *will* not) be conscious of his subjective weights, let alone the rationale — if there is one — for his assignment. But the empirical metatheorist can ascertain what these weights are by statistical study of scientific practice.

This brings us to a question about the case study method in metatheory. What can a single case study, say, of an episode in history of science, "prove"? It obviously cannot prove a statement of general form about how science proceeds, the Hasty Generalization Fallacy. It can, of course refute such generalizations, as is done in Carrier's paper. But when reading philosophers of science who rely on case studies, one notes that their favorite examples usually *refute generalizations that nobody made*. When Popper cites the quick *modus tollens* destruction of the Bohr-Kramers-Slater theory by the Bothe-Geiger experiment, what general statement does this instance falsify? It falsifies "No theorist ever abandoned his theory as a result of admitting a single clear refuter into the corpus." But who ever asserted that? You may say "Lakatos did". Not that I can find. Lakatos said a theorist need not do that, not that no one ever has. And that he need not was of course argued earlier by Quine and Duhem.[10] On the other side, Feyerabend is fond of Prout's hypothesis, as an example against Popper and Lakatos. But what generalization does the Prout episode refute? Something like, "No theory, having been long abandoned by almost everyone because of a clear falsifier [the atomic weight of chlorine = 35.5] has ever been revived on the basis of new evidence and theory." Did anyone ever assert that as a general principle, either descriptive or prescriptive? Not that I know of. Case studies are of intrinsic interest to the historian of science, being a legitimate scholarly activity (as, for political historians, "How did World War I break out?" is a worthwhile inquiry); and they can be a fruitful source of conjectures for the philosopher of

[10]But cf. A. Grünbaum, 'The Duhemian Argument', *Philosophy of Science* **27** (1960), 75–87.

science; but unless case studies are treated statistically, they are incapable of settling questions of metatheory, whether descriptive or prescriptive.

Since the correlation between theories' several distinct attributes (both internal and performative) and their verisimilitude cannot (at best) be other than stochastic, in order to test metatheoretical conjectures on a batch of case studies we need to analyze the historical data appropriately for statistical relations. *Thesis: Metatheoretical research should (1) make actuarial summaries of the properties and fates of scientific theories based on random sampling of episodes from the history of science and (2) apply formal analytic methods (e.g., psychometrics) to appraise metatheoretical conjectures*. Both descriptive and prescriptive generalizations should be tested in this way, rather than by the present informal, impressionistic (and often biased) reliance on selected case studies.

This radical and heretical thesis, which I find usually stuns people on first hearing, I label the Strong Actuarial Thesis. So far as I know, only two scholars (both, interestingly, clinical psychologists) have advocated the strong form explicitly, David Faust and myself.[11] But others have suggested doing

[11]D. Faust, *The Limits of Scientific Reasoning* (Minneapolis: University of Minnesota Press, 1984) was the first systematic defense and development of the idea that scientists should employ explicit statistical methods in appraising theories, sampling episodes from the history of science and making actuarial generalizations therefrom as to desirable properties of theories and strategies for concocting, appraising, and improving them. Faust relies on research by clinical, cognitive, social, and developmental psychologists to show that human cognitive processes are usually inefficient and that scientists are subject to the same deficiencies as others, although perhaps less so in some respects. See also Faust and Meehl, *op. cit*., note 9; Meehl, *op. cit*., note 8, pp. 10–20; Meehl, 'Appraising and Amending Theories: The Strategy of Lakatosian Defense and Two Principles that Warrant Using It', *Psychological Inquiry* **1** (1990), 108–141, and 'Meehl's Reply to the Commentators', pp. 173–180; Meehl, 'Subjectivity in Psychoanalytic Inference: The Nagging Persistence of Wilhelm Fliess's Achensee Question', in J. Earman (ed), *Minnesota Studies in the Philosophy of Science: Vol. X, Testing Scientific Theories* (Minneapolis: University of Minnesota Press, 1983), pp. 349–411, see pp. 371–372 (first presented at the Minnesota Center for Philosophy of Science, June 1980). In the latter, criticizing metatheorists' exclusive reliance on selected historical examples, I argued that assessing *metatheoretical* principles is "an inherently statistical problem, and that it cannot be settled except by the application of formal statistical methods". Faust and I are agreed that the considerations for appraisal of theory and metatheory are similar (if not identical), and it is obvious that an actuarial examination of *metatheoretical* principles will unavoidably statisticize *theories*' properties in relation to their performance. Starting in the other direction, the correlations of theory properties and scientist strategies with long term "theoretical success" as ascertained by actuarial study of scientific episodes, while initially *de*scriptive, becomes immediately *pre*scriptive in metatheory via the hypothetical imperative: "If you wish to succeed at the game of science, it usually pays, *ceteris paribus*, to do so-and-so." This advice is conjoined with the explanatory analysis: "Because [rational reconstruction]". For further discussion of this theory/metatheory relation see Meehl, 'Cliometric Metatheory: The Actuarial Approach to Empirical, History-Based Philosophy of Science', *Psychological Reports* **71** (1992), 339-467.

statistics on theory properties and performance.[12] Sulloway[13] also insists on actuarial studies, although he seems mainly interested in the *scientist's* demographic attributes rather than quantifying the *theory's* properties and performance. To my knowledge, the earliest published reference to such an idea appeared over a half-century ago in Reichenbach,[14] in his defense of the "identity conception" of probability (always a frequency) which, *prima facie*, appears not to fit how we appraise scientific theories.[15] But he does not exemplify or expand, and his student Salmon[16] never heard him do so orally. Thus we do not know whether the attributes he had in mind were "intrinsic" (e.g., mathematical, type of concept, formal relations among postulates) or "empirical" (e.g., predictive? specificity? number of replications? factual diversity?).

Despite its unusualness, the first part of the Actuarial Thesis seems, once stated, obvious and inescapable on the current view of metatheory as the (empirical) theory of scientific theorizing. Drawing lessons from the history of scientific successes *and failures*, the metatheorist aims to find out how (and why?) science "works", when it does. We know there is no guaranteed, automated truth-generating machinery, The Scientific Method, a set of strict rules. Rather we have a set of "principles", rough rules of thumb, guidelines, broad strategies, warnings and advice, "helpful hints", which we have come to believe *tend*, in an average sense and in the long run, to pay off. We know that each of these guidelines has been violated successfully, a historical fact that could easily be predicted from their mutual non-derivability (i.e., if principles P_1 and P_2 point in opposite directions as applied to a T, one of them will "win" in the long run and the other will "lose"). Since the whole enterprise is stochastic, we have no reason to be surprised at this, any more than we are astonished when a "reasonable decision" (bet on a good horse, buy life insurance, carry an umbrella) turns out, in the event, to be the "wrong" choice. Granting this for any proposed set of metatheoretical guidelines, how can the collection of scientific episodes provide guideline weights, qualifications, countervailings, interactions other than by statistical treatment? Suppose I think successful prediction of three novel facts corroborates a theory more than after-the-fact

[12]E.g., A. M. Diamond, Jr., 'The Polywater Episode and the Appraisal of Theories', in A. Donovan, L. Laudan, and R. Laudan (eds), *Scrutinizing Science: Empirical Studies of Scientific Change* (Boston: Kluwer Academic Publishers, 1988), pp. 181–198, see p. 181; P. Feyerabend, *Farewell to Reason* (London: Verso, 1987), see p. 171; L. Laudan, 'Normative Naturalism', *Philosophy of Science* **57** (1990), 44–59, see p. 46.

[13]F. J. Sulloway, 'Orthodoxy and Innovation in Science: The Influence of Birth Order in a Multivariate Context', paper presented at the meeting of the American Association for the Advancement of Science, New Orleans, LA, 16 February 1990.

[14]H. Reichenbach, *Experience and Prediction* (Chicago: The University of Chicago Press, 1938), see pp. 396–399.

[15]Cf. E. Nagel, 'Principles of the Theory of Probability', *International Encyclopedia of Unified Science* **1**(6) (1939), see pp. 65–66.

[16]Personal communication 15 August 1989.

explanation of six facts. You don't agree. Since we both admit that any such principle, if sound as a guideline, can at best be "correct" statistically, how could we possibly settle our dispute *except* by counting cases and computing summary statistics? So the first part of the Actuarial Thesis rests on the truism that *if a dispute is about a statistical relation, it can only be resolved by computing the statistics*. I have labeled this use of statistics *discriminative-validating*,[17] the question raised and answered being simply, "Is variable *X* related to variable *Y*, and in what form and degree?" Usually the only problematic features are those dealt with by statisticians (e.g., randomness of sample? bias of statistical estimator? best descriptive statistic? regression linear? adequate statistical power to detect a trend?).

The second half of the Actuarial Thesis is more controversial. While some statistics (chi-square, Pearson *r*, linear discriminant function) are almost entirely theory-free (except, of course, for the mathematical theory of probability), and in their pure discriminative-validating use one makes only those minimal inferences warranted by the statistician, other statistics are tied up with components of substantive theory, both as to their applicability and the content of inferences (causal, structural, latent entities) they are intended to provide. Factor analysis, path analysis,[18] and taxometrics[19] are methods employed to make theoretical inferences rather than merely for technological tasks of probabilistic prediction. Their value is currently not agreed upon, either by statisticians or life scientists, although it seems clear that conjectured causal interpretations of correlational data can sometimes be strongly *dis*corroborated by these methods. If one is a Popperian, that is presumably enough to warrant their use. Discussion of this use of statistics, which I have called the *structural-analytic* (better, I would now say, *causal-theoretical*) use, is beyond the scope of this paper. Suffice it to say that the empirical metatheorist should surely include them in the provisional set of analytic tools, and "by their fruits…."

I have elsewhere given an example of how statistics might be used in cliometric metatheory.[20] Briefly, one constructs an index of risky numerical prediction of an experimental measurement by combining (multiplicatively) narrowness of a theoretically tolerated interval *I* in ratio to an antecedent

[17]P. E. Meehl, *Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence* (Minneapolis: University of Minnesota Press, 1954), see pp. 11–15.

[18]J. C. Loehlin, *Latent Variable Models: An Introduction to Factor, Path, and Structural Analysis* (Hillsdale, NJ: Lawrence Erlbaum Associates, 1987).

[19]P. E. Meehl, 'Factors and Taxa, Traits and Types, Differences of Degree and Differences in Kind', *Journal of Personality* **60** (1992), 117-174. P. E. Meehl and R. Golden, 'Taxometric Methods', in P. Kendall and J. Butcher (eds), *Handbook of Research Methods in Clinical Psychology* (New York: Wiley, 1982), pp. 127–181.

[20]*Op. cit*., note 8; Meehl, 1990 and 1992, *op.cit.*, note 11.

range S (= Spielraum), with relative closeness of observed to predicted value (D = deviation), defining

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

and this index is then standardized appropriately. One also constructs an index of riskiness of curve type, an index of factual diversity, an index of reducibility ("up" and "down" in Comte's Pyramid of Sciences), etc. These crude indexes are then correlated, over a random class of theories, (a) with one another and (b) with theories' long-term fates. A theory that has been treated in textbooks for, say, 50 years as "established" and is no longer a focus of experimental test, we count as having high verisimilitude if we are scientific realists. Instrumentalists and fictionists will avoid that locution, saying (more safely) that they confidently expect T to do well in the future, extrapolating from its past performance. For the latter group, only the discriminative-validating use of statistics is involved; for realists, part of the use is structural-analytic, and hence, more problematic.

Verisimilitude is a matter of degree, some theories having more "truth-likeness" than others. It is unfortunate that logicians' attempts at rigorous, general definition have thus far failed to command consensus. I believe that the concept is essential to metatheory, that no one familiar with the theorizing of working scientists could conceive a metatheoretical account of science lacking some such notion. My conjecture is that philosophers have gone about defining it in the wrong way (e.g., in terms of infinite consequence-classes), presupposing the existence of some *single unidimensional property* which we have not as yet been clever enough to ferret out. I have suggested a different way of looking at verisimilitude[21] that stays closer to the kinds of questions asked by the scientist in appraising the truth-likeness of a theory. The similitude of two theories is the extent to which their postulates agree at Levels I–X in Table 1.[22] The verisimilitude of a theory is the similitude to Omniscient Jones's theory T_{OJ}. A postulate cannot satisfy a level (k + 1) in the list if it has failed level k, i.e., the levels are what psychometricians call "Guttman-scalable". A crude verisimilitude index could be defined for a theory as the mean of the levels passed by its n postulates.

Of course we do not have access to Omniscient Jones's T_{OJ}. Lacking it, we identify the class of theories that have been ensconced unchallenged in textbooks and standard treatises for, say, the last 50 years, and treat each of these as the correct, true theory of its domain. We then examine its former

[21] *Op. cit.*, note 8.

[22] An earlier version of this table was published in Meehl, 1990, *op. cit.*, note 9.

Table 1

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

competitor theories, including earlier forms of itself, with respect to the ten similitude levels. At various stages of scientific development (e.g., a theory's "half-life", based on number of experiments rather than time) it will have such-and-such a verisimilitude score. Over a sample of theories, various aspects of theory performance, such as the corroboration index C_i explained above, can be correlated with verisimilitude. (The weights assigned to the ten levels can be chosen in several ways, the important point there being that it does not matter much, a well known theorem in classical psychometrics.[23])

Philosophers may be troubled by the circularity of all this. It has two components. First, even discriminative-validating statistics begin by sorting historical theories into the two boxes, "successful" and "abandoned". But not all of them will stay put, witness Newton and Prout. Doesn't this invalidate the whole procedure? This is a good question, but it does not distress psychologists at all. We are case-hardened to reliance on "fallible criteria" (a phrase that doubtless sounds like an oxymoron to logicians). Only rarely does the clinical psychologist have access to what is termed a Gold Standard Criterion, one that is perfect in classifying, ranking, or measuring individual patients because it is *definitional* of the attribute, or has been empirically shown to be quasi-perfectly correlated with the definitional one. For example, in building a mental test for assessing schizophrenia, we start with the psychiatrist's chart diagnosis, although we know (from statistical studies) that it is somewhat *unreliable* and, hence, *imperfectly valid*. So long as this official diagnosis possesses some validity, it is easy to show mathematically that we can identify verbal items valid for schizophrenia, and that a composite score obtained by tallying the count on a batch of such items (*score* on a multi-item psychometric *scale* or *test*) can easily be more valid that the original diagnostic criterion. This "bootstraps effect" in psychometrics was so christened by Cronbach and

[23]Meehl, 1992 *op. cit*., note 11.

myself;[24] and we continue to use the term despite its subsequent introduction for a different (but related) concept by Glymour, who informs me that three other disciplines (chemistry, geology, statistics) also employ the word, the contexts being different but analogous. In the actuarial study of theories, it does not matter that a few theories will some day be re-classified in our accepted/discarded dichotomy, since the theory properties are only related probabilistically to this criterion anyway.[25]

The more serious threat of circularity is that, aside from a few ultimately reclassified cases, the *list of attributes* used by textbook writers largely coincides with the list we are correlating ("validating") with the textbook fate of theories. Isn't this viciously circular? No, because *the correlations are empirical findings* — they are not forced to show statistical coherence, as a mere matter of definition. Our line of reasoning is like that involved in constructing an omnibus intelligence test. We start with tasks commonsensically classified as "cognitive" in nature. The statistics take over from there, and we find that there is a big general (mathematical) factor underlying the intercorrelation matrix, that this factor steadily develops with chronological age up to adulthood, that its growth is impaired by damage to the brain, that the score correlates with ratings by teachers and peers, that it correlates with performance in certain experimental tasks, that it predicts academic and vocational performance (even who gets into *Who's Who*!), that it is rather strongly heritable, and so on. This complicated network of relations comes to constitute our "implicit definition" of the theoretical construct *g*, and we do not worry about whether it offends some people to call it "general intelligence". This process is psychometric bootstrapsing in a more interesting sense than merely finding items that correlate with a crude, pre-analytic "criterion", although that initial stage remains a crucial part of the whole research program. In bootstrapsing metatheory, an important task on the *conceptual* side is to show why one would expect a strong correlation between verisimilitude and a composite index of theory performance. For the crude index "experiments successfully predicted", I have offered such a proof.[26]

I hope it is not indulging in an *ad hominem* to alert the reader that many are tempted to dismiss the actuarial approach out of hand. I can mention three areas where statistical methods of interpreting data are strongly resisted, namely, clinical judgment, biological classification, and general history. As to the first of these, there is a sizable body of empirical research (some 175 studies) showing that the conventional informal, impressionistic, in-the-head method of diagnosing and prognosing is either equal to or significantly less

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

[25]*Op. cit*., note 14, pp. 307–308, 331–333.

[26]*Op. cit*., note 8.

accurate than the formal method, applying an algorithm to the data.[27] I know of no long-standing controversy in social science on which the empirical evidence is so massive, qualitatively diverse, and consistent.

In biological classification, the value of formal taxometric methods[28] is hotly disputed. "Traditional" taxonomists have sometimes reacted to the numerical approach not with what Freud called "benevolent skepticism" but with dogmatic rejection and hostility.

In (general) history, a similar intense conflict has taken place, quantitative historians having founded their own journal and professional society; and here also the methodological issue remains unsettled.[29] I do not have expertise to form an opinion as to these latter two fields, but I am probably the world authority on the first. I mention them here as evidence that — for whatever

[27]H. R. Arkes and K. R. Hammond, *Judgment and Decision Making: An Interdisciplinary Reader* (New York: Cambridge University Press, 1986); R. M. Dawes, 'Probabilistic versus Causal Thinking', in W. M. Grove and D. Cicchetti (eds), *Thinking Clearly About Psychology. Vol. 1: Matters of Public Interest* (Minneapolis: University of Minnesota Press, 1991), pp. 235–264; R. M. Dawes, D. Faust, and P. E. Meehl, 'Clinical versus Actuarial Judgment', *Science* **243** (1989), 1668–1674; Dawes, Faust, and Meehl, 'Statistical Prediction vs. Clinical Prediction: Improving What Works', in G. Keren and C. Lewis (eds), *Methodology and Quantitative Issues in the Analysis of Psychological Data* (Hillsdale, NJ: Lawrence Erlbaum, 1993), pp. 351-367; D. Faust, 'What If We Had Listened? Present Reflections on Altered Pasts', in Grove and Cicchetti (eds), *ibid*., pp. 185–216; L. R. Goldberg, 'Human Mind versus Regression Equation: Five Contrasts', in Grove and Cicchetti (eds), *ibid*., pp. 173–184; H. G. Gough, 'Clinical versus Statistical Prediction in Psychology', in L. Postman (ed), *Psychology in the Making* (New York: Knopf, 1962), pp. 526–584; W. M. Grove, 'Clinical Inference from Psychological Tests: Last Nails in the Coffin', paper presented at the Minnesota Psychological Association, Minneapolis, May 1986, and personal communication 1991; B. Kleinmuntz, 'Recent Developments in Computerized Clinical Judgment', in Grove and Cicchetti (eds), *ibid*., pp. 217–234; *op. cit*., note 17; P. E. Meehl, 'What Can the Clinician Do Well?' in D. N. Jackson & S. Messick (eds), *Problems in Human Assessment* (New York: McGraw-Hill, 1967), pp. 594–599 (Reprinted in Meehl, *op. cit*., note 24, pp. 165–173); Meehl, 'Causes and Effects of my Disturbing Little Book', *Journal of Personality Assessment* **50** (1986), 370–375; J. Sawyer, 'Measurement *and* Prediction, Clinical *and* Statistical', *Psychological Bulletin* **66** (1966), 178–200; J. O. Sines, 'Actuarial versus Clinical Prediction in Psychopathology', *British Journal of Psychiatry* **116** (1970), 129–144; but cf. R. R. Holt, *Methods in Clinical Psychology*, vol. 1, 2 (New York: Plenum, 1978).

[28]D. L. Hull, *Science as a Process* (Chicago: University of Chicago Press, 1988); N. Jardine, and R. Sibson, *Mathematical Taxonomy* (London: Wiley, 1971); P. H. A. Sneath, and R. R. Sokal, *Numerical Taxonomy* (San Francisco: Freeman, 1973).

[29]W. O. Aydelotte, A. G. Bogue, and R. W. Fogel (eds), *The Dimensions of Quantitative Research in History* (Princeton, NJ: Princeton University Press, 1972); J. Barzun, *Clio and the Doctors: Psycho-History, Quanto-History, and History* (Chicago: University of Chicago Press, 1974); A. G. Bogue, *Clio and the Bitch Goddess: Quantification in American Political History* (Beverly Hills, CA: Sage Publications, 1983); R. Floud, 'Quantitative History and People's History: Two Methods in Conflict?' *Social Science History* **8** (1984), 151–168; R. W. Fogel, and G. R. Elton, *Which Road to the Past? Two Views of History* (New Haven, CT: Yale University Press, 1983); G. Himmelfarb, *The New History and the Old* (Cambridge, MA: Belknap Press of Harvard University Press, 1987); C. Tilly, 'The Old New Social History and the New Old Social History', *Review* **7** (1984), 363–406; K. W. Wachter, E. A. Hammel, and P. Laslett, *Statistical Studies of Historical Social Structure* (New York: Academic Press, 1978); further references and discussion can be found in Meehl, 1992, *op. cit*., note 11.

reason(s) — any proposal to formalize and numerify the assessment of large and complex data sets (instead of proceeding in the conventional informal, impressionistic way) tends to arouse resentment and anxiety.[30]

In current practice the properties of theories, whether conceptual or empirical, are assessed in a subjective, impressionistic, informal manner. Scientists and metatheorists report their *judgments*, like clinicians who invoke their "clinical skill" in diagnosing patients, relying on the purported (but rarely checked) inferential powers conferred by years of clinical experience. If an actuarial approach to theory evaluation and validation of metatheory were adopted, there would still be a place for the exercise of scholarly judgment. Some attributes of theories will resist objective numerification for a long time. For example, although Popperian risky test or Salmonean damn strange coincidence can sometimes be formulated in terms of a narrow tolerance of observational value in relation to an antecedent atheoretical Spielraum, there are surprising facts that are not of the point-prediction sort (e. g., animal learning facilitated by mildly punishing correct responses; improved recall of nonsense syllables after a short time lapse [the "reminiscence effect"]; the bright spot effect in Fresnel's experiment). Conceptual intrinsic properties such as depth, elegance, richness, beauty are sometimes weighted heavily by physicists, and such subtle attributes are matters of judgment. Even here, however, the psychometrician has a contribution to make. There is a vast body of empirical research (and lore) concerning human judgments, and psychologists know a lot about them. For example, industrial psychologists have perforce relied on rating scales as a cheap, convenient, "natural" (but still

[30]I do not include here controversy about psychologists' use of *meta-analysis* to study interventions because (a) the objectors do not seem distressed or biased and (b) the objections are mainly to misapplication rather than to the idea as such. In meta-analysis the numerous empirical studies of a pragmatic intervention (e.g., efficacy of psychotherapy) are summarized in formal statistics relating the targeted change ("effect size") to various independent variables characterizing the studies. *The research study is the statistical unit*, being dealt with as individual patients, students, or subjects conventionally are *within* studies. Meta-analysis was devised for evaluating a large mass of studies varying in design and parameters and, hence, in outcome (G. V. Glass, B. McGaw, and M. L. Smith, *Meta–Analysis in Social Research*, Beverly Hills, CA: Sage, 1981; J. E. Hunter, F. L. Schmidt, and G. B. Jackson, *Meta–Analysis: Cumulating Research Findings Across Studies*, Studying Organizations: Innovations in Methodology, vol. 4 Beverly Hills, CA: Sage, 1982). Its original aim was *technological*, to aid in making policy judgments, but it is increasingly used to *appraise theories*, and for that it is inappropriate (P. E. Meehl, 'Why Summaries of Research on Psychological Theories are Often Uninterpretable', *Psychological Reports* **66** (1990), 195–244 (also in R. E. Snow and D. Wiley (eds), *Improving Inquiry in Social Science: A Volume in Honor of Lee J. Cronbach* (Hillsdale, NJ: Lawrence Erlbaum Associates, 1991), pp. 13–59); S. L. Chow, 'Meta-analysis of Pragmatic and Theoretical Research', *Journal of Philosophy* **121** (1987), 259–271). But meta-analysis does have a deep affinity with quantitative methods used in these three areas, and with the Faust-Meehl Actuarial Thesis. Their shared core idea is that the human mind, faced with the task of informally combining diverse complex and countervailing data related stochastically to a criterion, is a less effective information processor than commonly supposed; and that even a less-than-optimizing formal procedure, operating on actuarial derived values, will often do a better job.

quantifiable) source of information about personal traits, efficiency, managerial skill, and even physical "work products". Rules of thumb for constructing such scales are well known and utilized by all competent psychologists.[31] It is also known that in many domains human judgments behave statistically rather like single mental test items, so that the reliability of a pooled judgment can be quite accurately predicted from the pairwise correlations between judges (utilizing the Spearman-Brown prophecy formula). Thus if we found that pairs of scientists correlated only $r = 0.60$ (a typical value for subjective ratings of personal attributes like "intelligence" or "shyness"), we could obtain the average of seven judges to achieve a reliability of $r = 0.91$. Important general point: *The unavoidability of employing subjective judgments by skilled persons does not imply that statistical methods are inapplicable. On the contrary, the most efficient use of such judgments is usually* — I do not say always — *achieved by subjecting them to psychometric treatment*.[32]

Because Carrier's paper deals with realism, and it is the metatheoretical position I espouse, I have argued for the Actuarial Thesis within a realist framework. Despite the present unsatisfactory state of the verisimilitude concept,[33] I remain convinced that it is an indispensable notion in any rational reconstruction of scientific theorizing. My main reason for that conviction is that working scientists in every field I know about regularly invoke it, whether or not they use the *term* 'verisimilitude' or are interested in philosophy of science. So my response to the philosophers is, "If you haven't explicated closeness-to-the-truth satisfactorily, I urge you to continue working on it." However, it should be clear that the Actuarial Thesis does not depend, even slightly, on the verisimilitude concept, or on scientific realism generally. A consistent positivist, phenomenalist (if any are left!), fictionist, or instrumentalist can employ actuarial summaries and psychometric analyses of scientific episodes, and *should* do so, for the same reason that a realist should, i.e., the cognitive limitations of the human mind in processing information about complex stochastic relationships. The instrumentalist will correlate "intrinsic" (formal and conceptual) theory properties with "empirical" (empirical performance, psychosocial) properties and relations,[34] none of these correlations involving the idea of verisimilitude. Forecasting future instrumental success from track record to date, plotting graphs of progressive and degenerating

[31]See, e.g., J. P. Guilford, *Psychometric Methods* (New York: McGraw-Hill, 1954).

[32]An interesting historical sidelight. The inception of psychology as a quantitative experimental science is often traced to Bessel's (1823) "personal equation" in astronomy calibrating observers of star transit times and correcting for individual bias. Bessel's research arose out of the famous 1796 episode of Royal Astronomer Maskelyne discharging assistant Kinnebrook because the latter's transit readings were "wrong," i.e., deviated systematically from Maskelyne's (E. G. Boring, *A History of Experimental Psychology* (New York: Appleton-Century, 1929)).

[33]E.g., rejected by such an incisive and deep Popperian thinker as J. W. N. Watkins, *Science and Scepticism* (Princeton, NJ: Princeton University Press, 1984).

[34]Faust and Meehl, *op. cit*., note 9.

programs, choosing directions of further experimentation, relating classes of concepts to predictive success, selecting which portions of the nomological network to amend, deciding whether to conduct a Lakatosian defense,[35] making technological extrapolations — all these can be conducted in the light of cliometric analysis, within a non-realist framework. The situation is closely analogous to that of an industrial psychologist studying psychological tests, interviews, job performance, labor turnover, factory morale, accident rate, and the like, but disclaiming interest in the "reality" of mental dimensions putatively revealed by factor analysis,[36] causal connections corroborated by path analysis,[37] or latent taxa detected by taxometrics.[38]

A philosopher may resist the Strong Actuarial Thesis, misconceiving it as an attempt to replace philosophy of science by cognitive psychology and sociology of knowledge, conflating the contexts of discovery and justification, fallaciously reducing the prescriptive (normative) to the descriptive (factual), abandoning the aim of rational reconstruction — in effect, liquidating the distinctively philosophical enterprise. While some scholars entertain such ideas (if I understand them correctly), let me dissociate myself from them in the strongest and clearest possible terms. I view metatheory as the rational reconstruction of scientific knowledge, as the scientific theory of scientific theories. Like other scientific theories, it aims to *explain* the phenomena of its domain. We want to *understand* why science "works", not merely to ascertain the empirical correlates of its working well or badly. Such correlations are the first-order inferences of the empirical domain: scientific change. I conjecture that a satisfactory explanation of scientific change will include components of "rationality". There is no more reason for a metatheorist to deprive himself of logic and mathematics (especially probability theory) as part of his conceptual toolkit than there would be for an economist studying the success and failure of business firms (statistically!) to disallow all references to irrational executive decisions. The relation of prescription to description in characterizing science is beyond the scope of this paper, but I have presented an elaboration and defense of the above remarks elsewhere.[39] Suffice it to say here that metatheoretical prescriptions and proscriptions are best viewed as hypothetical imperatives, based on empirical generalizations from history of science: "If you wish to have scientific success, the odds are better if you do so-and-so". That advice is, of course, already a piece of instrumental rationality. But the metatheorist, like any other theorist, asks *why* the principles work. Answering that question

[35]Meehl, 1990, *op. cit*., note 11.

[36]P. E. Meehl, 'Four Queries About Factor Reality', *History and Philosophy of Psychology Bulletin* **5** (no. 2) (1993), 4-5.

[37]*Op. cit*., note 18.

[38]*Op. cit*., note 19.

[39] Meehl, 1992, *op. cit*., note 11.

will, I am confident, involve the kinds of analysis that philosophers have traditionally engaged in, including some "armchair epistemology."

Since nobody claims that scientists appraise theories optimally, nor that philosophers of science have now settled the major metatheoretical issues (by a combination of armchair epistemology and case studies), it is reasonable to hold that the Actuarial Thesis is a live option, and that empirical research should be conducted on it.