Behavior Therapy 23, 195-211, 1992

Using Scientific Methods to Resolve Questions in the History and Philosophy of Science: Some Illustrations

DAVID FAUST

University of Rhode Island

PAUL E. MEEHL

University of Minnesota

The factors that account for successful scientific efforts are disputed among philosophers and historians of science, who rely heavily on impressionistic and case study methods to support their positions. Resolution of many such debates will require more adequate methods for sampling, analyzing, and integrating the historical track record. A more accurate description and understanding of scientific processes and successes, and the development of decision aids for such higher level scientific judgments as theory appraisal, will offer practical help to the working scientist.

The primary aim of this special issue, as we understand it, is to get behavioral scientists to think more about the philosophical underpinnings of their trade, or to think more like philosophers. We will contend, however, that the philosophers to whom they turn and those in related disciplines (sociology, history, and psychology) whose specialty is the study of science, should think and act more like scientists. As we will argue, greater rigor in the study of science can, and someday will, revolutionize the philosophy of science and deliver practical goods to the working scientist.

Jones' Amazing Dissertation

Imagine a planet where individuals live a very long life. One of the members of this world, Jones, is laboring on his dissertation. Jones has invented some rules of thumb for acquiring knowledge, which he has labeled the scientific method. These guides address such matters as systematic observation and recording, and strategies for countering biased perceptions. Jones presents this material to his dissertation committee and they, being demanding as most committees are, find it insufficient and require a test of Jones' ideas. The committee, unlike most, is able to agree

We would like to thank Drs. Robyn Dawes, Steve Fuller, Lee Sechrest, and Al Silverstein for reviewing an earlier draft of this manuscript and offering helpful suggestions. We would also like to thank Dr. Leslie Yonce for her help in preparation of the manuscript. Requests for reprints should be addressed to: David Faust, Department of Psychology, Chafee Center, University of Rhode Island, Kingston, RI 02881.

about something, in this case certain rules for appraising the success of Jones' method. Most important in keeping score are whether the method advances the understanding of causal mechanisms not currently understood (we will assume there are reasonably well worked out procedures for such determinations) and whether it permits the prediction of new phenomena. The question is how to carry out these tests.

The committee "suggests" a plan. Jones is to travel to the planet Earth and plant this so-called scientific method in the mind of some intellectual luminary, which he will take as his own and teach to others. These individuals will then apply the method to see how well it works. Success of the method will be compared to the cognitive achievements of other human enterprises that purport to acquire or represent knowledge, such as politics and religion. But what luminary to seek out? Jones considers a William Shakespeare, but Jones' Chair questions the company this choice keeps, so Francis Bacon becomes the one.

Once imparting this knowledge, or these basic guides or procedures, Jones departs the scene to take care of his neglected personal life and to await the results of his project. During the 350 years or so he is away, things get somewhat out of hand; and by the time he returns, thousands upon thousands of individuals, who are now called scientists, have applied the method, but often in varying ways and with varying subject matters. There have now been hundreds of thousands of applications of the method, and it has been substantially altered and elaborated upon, with technologies developed that extend the capabilities of the naked senses or the mind, such as telescopes, microscopes, accelerators, and computers. Working from this gigantic and diverse data base, Jones' task is to make sense of these hundreds of thousands of applications, or studies, of the scientific method. Jones' task is two-fold. First, he must compare the overall success of the method to that of alternative systems. Second, as outcomes are far from uniform, he must distill the relations between variations in method and outcome.

The first task is easy. Even casual observation shows that the scientific method is generally superior to other systems in achieving its cognitive aims or advancing its knowledge base. Whereas the politicians and religious leaders seem to be arguing over the same basic matters and presenting the same kinds of evidence their predecessors did over 300 years ago, the dialogue of the scientists has advanced remarkably. And the scientists are now able to predict a wide variety of things, often with impressive precision, such as the occurrence of eclipses, the effects of various disease processes, and the speed of a ball on an inclined plane. In contrast, the political and religious leaders are no better able to predict than they were before.

It is also obvious, however, that despite this overall superiority, the effectiveness of scientific endeavors varies tremendously. Some scientific fields have progressed much more rapidly than others, and some fields, or specialties, have failed completely. Some problems have yielded quickly and some not at all. Some conclusions or beliefs have been warranted and others seem to have been badly misplaced. Many different methods have been employed and some have succeeded better than others, but even so, in some cases methods that had been generally successful led to failure and those with a much lower ratio of successes sometimes led to stunning achievements. Given the mass of data, the less than complete (and not necessarily accurate) historical record, and variation in outcome, the second part of Jones' task is daunting.

The Study of Science

If one views applications of the scientific method since its inception as simultaneously providing tests of the topic the investigator had in mind and tests of the scientific method itself, then one attempting to represent or discern this mass of data accurately—to describe and explain the success of science generally and specifically—faces precisely the task of our hypothetical Jones. And, indeed, this is the task confronted by those who attempt to describe, or reconstruct, the history and success of science. How might one go about such a difficult task, and why might one bother to make the effort?

Methods for Studying Science

One option is to use subjective, impressionistic judgment to evaluate the data. (In this context, what we mean by data is the information available on scientific occurrences, as is found in textbooks, historical accounts in books and journals, scientists' work records and notes, and other material of that sort.) Impressionistic judgment has an important place in such an enterprise. For example, until a better alternative is developed, subjective judgment is the best source for detecting possible predictive variables and for observing certain types of things. No one has yet invented a computer that can read through Darwin's notebooks and draw conjectures about the factors that may have shaped Darwin's thinking.

Subjective, impressionistic judgment, however, has marked limitations, particularly in the context of hypothesis testing or justification. For example, observation and impression are prone to the operation of various biases (Arkes, 1981; Dawes, 1988; Faust, 1984, 1986; Kahneman, Slovic, & Tversky, 1982; Meehl, 1973; Nisbett & Ross, 1980). Dramatic occurrences may be given more attention, or unduly influence judgment, in comparison to less dramatic but informative data; preliminary hypotheses have a decided advantage in the hypothesis-testing process; and false associations are easily formed between variables. Thus, for example, when analyzing this giant data base, one might draw overly broad generalizations from dramatic discoveries, selectively seek out instances that support one's presuppositions about the historical track record, or link certain methodological procedures or preferences with success when, in fact, these procedures produce poor outcomes just as often or are inferior to other alternatives.

Perhaps a more important problem is the limited capacity of the human mind to integrate complex and diverse data. The human mind is not adept at weighting variables optimally or deciphering complex configural relations among data. The huge body of research on clinical versus actuarial judgment convincingly shows that even crude, non-optimized decision procedures that combine information in a linear manner consistently equal or exceed the accuracy of human judges (Dawes, Faust, & Meehl, 1989; Meehl, 1954).

A related lesson (that should not be lost on one who studies science and that applies to the efforts of such an inquirer as well) is that science succeeds in large part to the extent it goes beyond impressionistic judgment. The scientific enterprise shares a subjective component with other human endeavors, but one point of demarcation is the use of methods that supplement, exceed, or transcend the unaided mind. Each time the scientist *counts*, he or she tacitly or explicitly acknowledges the limitations of impressionistic judgment, as he or she does when, for example, data are represented graphically, relations are expressed in mathematical terms, or instruments are substituted for the eye or ear. As Meehl (1986) has stated:

> Surely we all know that the human brain is poor at weighting and computing. When you check out at a supermarket, you don't eyeball the heap of purchases and say to the clerk, "Well it looks to me as if it's about $17.00 worth; what do you think?" The clerk adds it up. There are no strong arguments from the armchair or from empirical studies… for believing that human beings can assign optimal weights in equations subjectively or that they apply their own weights consistently. (p. 372)

To which Dawes, Faust, and Meehl (1989) have added:

> It might be objected that this analogy, offered not probatively but pedagogically, presupposes an additive model that a proponent of configural judgment will not accept. Suppose instead that the supermarket pricing rule were, "Whenever both beef and fresh vegetables are involved, multiply the logarithm of 0.78 of the meat price by the square root of twice the vegetable price"; would the clerk and customer eyeball that any better? Worse, almost certainly. When human judges perform poorly at estimating and applying the parameters of a simple or component mathematical function, they should not be expected to do better when required to weigh a complex composite of those variables. (p. 1672)

A related approach is to perform case studies, that is, intensive analysis of episodes in the history of science or of scientists. One might examine how members of the scientific community responded to one or a series of theories or disputes, or might trace the course of one or more scientists over time, or over their careers. There are sound reasons to attempt such efforts (e.g., opportunity for in-depth study, intrinsic interest), and certainly in the context of discovery (Reichenbach, 1938) these efforts can be particularly worthwhile and lead to conjectures that ultimately gain support. However, this is a poor method of corroboration. In testing ideas or possibilities about the history of science, or the scientific method, case study works best where it is not needed, which is to disconfirm extreme (or absurd?) statements of the type that leading historians or philosophers of science most usually utter only in caricature (i.e, when misrepresented by others). For example, had Popper claimed that no theory has ever been successfully or correctly retained despite apparent falsification, then a single case study that convincingly showed otherwise might be decisive. However, neither Popper nor Lakatos nor other leading philosophers have said such things. Even if they had, and even were such strong programs disconfirmed, one is left with the relatively uninformative conclusion that the frequency of contrary occurrences is greater than zero, but how much more—less than 5%, more than 50%, more than 95%—we do not know. It also follows that if various historians or philosophers made absolute but conflicting claims, all but one of them (and possibly all of them) must be wrong. Point: In this context, case studies inform us about a position no one is likely to have put forth, or tell us that things are less than absolute, a proposition that rarely is in doubt and disproof of which reduces the remaining level of uncertainty microscopically. That such "hypothesis testing" is seen as informative speaks to the lack of background knowledge.

The existence of exceptions to virtually any methodological dictate or advice one might offer reflects a basic characteristic of scientific processes—they are stochastic (probabilistic). Successful scientists can vary tremendously in what they have done, successful and unsuccessful scientists may pursue the same methods, and the same scientist may succeed one time but not another. This world is stochastic not only in the linkages between scientific methods or strategies and outcome, but also in the connections between facts and theories, for the same set of facts can be covered with comparable plausibility by alternate theories. In a heterogeneous and probabilistic world, one attempting to test impressions about scientific processes must be deeply worried about representative sampling.

The representativeness of the case study method is limited not only by the typically low sample size. More importantly, the cases are not selected randomly but rather to illustrate (prove?) historians' or philosophers' particular impressions/reconstructions of science. For example, Popper has not conducted *random* selection to test his views of falsification, but rather has cited episodes in the history of science to *illustrate* his program. However, the diverse data set makes such efforts likely to produce supportive instances, even should these features be rare or nonrepresentative of the scientific enterprise as a whole. One can acknowledge the important contributions of impressionistic methods and case study and simultaneously argue that they are severely limited tools by themselves for deriving an accurate description of science past and present (Donovan, Laudan, & Laudan, 1988; Laudan et al., 1986). Where there is a *massive* heterogeneous and probabilistic data base, impressionistic and case study methods will not do the job; more powerful analytic tools are required, such as representative sampling and decision aids for organizing and representing the available data. This need for more refined methods of study has been almost completely disregarded among those involved in the study of science, including philosophers, historians, sociologists, and even psychologists, who should know they are not exempt from the very factors and limits they themselves have identified as necessitating scientific method.

Rationale for Studying Science

One may accept the need to apply more rigorous methods to the study of science but still question the point of the whole enterprise, or of the attempt to derive an adequate description of scientific episodes and the factors underlying success. One response is that this is one of the great intellectual questions. Determining the keys to scientific success has enormous implications for epistemology, or for those interested in methods for acquiring knowledge. To the reader who believes that this puzzle is already solved—that philosophers or historians *or scientists* clearly understand what makes science succeed—we would suggest that the standard textbook descriptions are far from adequate. If brilliant minds in philosophy and science have differed sharply on the underlying components of success, if scientific practices are diverse and probabilistically related to outcome, if no adequate description of the scientific enterprise or historical episodes is currently available, and if impressionistic judgment is inadequate for uncovering answers, there is little reason for smug confidence.

Of greater appeal to the pragmatically minded, an adequate or improved description of science should provide guidance for the working scientist. In the hundreds of thousands of applications of scientific method, the scientist has done something and something has resulted. As noted, the relations between the scientist's actions and outcome are probabilistic For example, Smith does A which results in X, and Jones does B which also results in X. Alternatively, Smith does A which results in X, and Jones does A which results in Y. Given the stochastic nature of this process, the scientist must operate in a context of uncertainty and thus is like the gambler placing bets. (Note that our context here is epistemological, not ontological—we are addressing methods for acquiring knowledge, not the nature of knowledge.) Decisions to accept a theory, to modify it, to modify it in one way versus another, or to abandon the theory entirely are choices that are rarely, if ever, made with absolute certainty. For both the big and little determinations, one is constantly asking, "Is this the best move to make?" One way to decrease the uncertainty of such decision making is to access that portion of the empirical track record that entails comparable situations and to study the outcome of different moves. Of course, this is often done subjectively, for example, "Maybe I'd better replicate first before publishing because in situations like this (surprising results, a temperamental piece of equipment, data that somehow do not seem right, or whatever), it so often happens that…." However, as we have argued, subjective impression is a poor way to determine relative frequencies or representativeness, and a more rigorously and accurately derived compilation is needed. Analysis of the track record might show, for example, all other things being equal, that the various moves being pondered show meaningful differences in success ratio (e.g., are usually versus occasionally successful; or, perhaps, are almost never successful).

As with human behavior, in science past track record or payoff may often be the best predictor of future payoff. Certainly it is not far-fetched to suggest that what has succeeded more often before is more likely to succeed than what has succeeded less often. To assume otherwise suggests that a scientist who completely ignores the past track record (e.g., the benefits that control groups commonly provide) is behaving rationally. Nothing in this assumes that the probabilistic relations between action and outcome cannot or will not change. The problem, however, is determining whether such changes have taken place and what this means pragmatically in the context of decision making. This only illustrates a point we readily acknowledge—there will be exceptions to relations found in the historical track record (which of course is the same as saying the relations are probabilistic), and as such this record will not provide a set of unassailable rules or absolutes but rather guidelines or rules of thumb.

A Scientifically Based Metatheory

If knowing what worked before helps to determine what is likely to work later, then we need some systematic method for compiling the historical track record or integrating data on scientific episodes. We will suggest the general outlines of such a program, starting with working assumptions and ending with illustrations of research topics and methods of study.

Working Assumptions

The necessity of presenting working assumptions in condensed form risks an unintended tone of dogmatism. The reader who desires greater detail can consult various sources (Faust, 1984, in preparation; Meehl, 1990a, 1990b, 1990c, 1992a, [2002, 2004]).

1. *Science is the best available method for acquiring knowledge.*

(Jones has already established this point).

2. *Best available and optimal should not be conflated.*

Knowing that science is the best game going and has produced remarkable achievements establishes little about its approximation to the optimal. A sprinter may leave other runners in the dust, but from this we would not conclude that she is nearly as fast as a jet plane, can levitate across water so that boats are not needed, or that there is no reason to pursue new means of transportation. One might consider the evolution of the scientific method, especially the development of various decision aids and measuring instruments that extend the power of human senses or cerebration. However, scientific method provides little aid where it is perhaps most needed: in the integration of data.

Scientific method and the efforts of the scientific community help to counter certain frailties of the human mind, such as the tendency to overattend to positive instances to the neglect of negative ones. Overattending to positive instances leads to false belief in associations among variables or an overestimation of associative strength, a failing that can be corrected through such basic procedures as calculating a correlation coefficient or constructing a co-variation table.

Thus, if the clinical psychologist has come to believe that accented eyes in human figure drawings are associated with paranoid tendencies, a simple tally may show that the two sometimes do co-occur, but no more frequently with the condition of interest than with other conditions. In contrast, the scientific method provides virtually no guidance for integrating results across multiple experiments or studies *to test scientific theories*. In fact, until recently, there were few systematic approaches for combining data across studies in order to address far more rudimentary questions than the viability of a theory (e.g., such questions as the effect of an intervention in psychotherapy). Meta-analysis has provided an important tool for testing such basic questions, and it does offer one means for assisting the human mind in integrating data, but it is *not* a tool for theory evaluation of the sort we are advocating (see below).

3. *The scientific game is a stochastic or probabilistic one.*

Chance and uncertainty are inherent parts of the scientific process. A scientist may violate virtually every norm and succeed, and another may make all the right moves and fail. The same procedures may lead to different outcomes, and different procedures may lead to the same outcome. This is *not* to argue that any move the scientist makes is a blind guess or a throw of the dice and hence anything goes. For example, it would be absurd to maintain that adding numbers incorrectly is every bit as good as getting them right. In fact, we are arguing almost the opposite, that the likelihood of success varies depending on the moves that are made, and that some moves are more likely, or much more likely, to produce success than others. Nevertheless, to a varying extent, the relationships remain uncertain.

A central aim is to minimize or decrease the uncertainty of decision making, or to reduce the element of chance. For now, however, and perhaps forever, there will be a lack of certitude in scientific decision making. Investigation of the historical track record will not allow us to develop absolute rules of procedure but rather generalizations or principles that offer informed guidance.

4. *The human mind is a limited instrument for integrating data.*

Both authors share the seemingly paradoxical view that the psychology of science will make its most positive contribution by detailing human cognitive limits, thereby signaling and outlining needed forms of decision aids or corrective mechanisms that eventually will be incorporated into scientific method. There is a massive body of data demonstrating the limits of human integrative abilities. For example, the 100-plus studies on clinical versus actuarial judgment show that for purposes of data combination, the accuracy of simple methods, such as those that depend on little more than counting, almost always equal or exceed that of subjective or unaided human judgment (see Dawes, Faust, & Meehl, 1989; Meehl, 1954). Other research demonstrates individuals' difficulties combining increasing amounts of information, deciphering interactions or interrelations between multiple cues or variables, and analyzing complex or configural relations (see Faust, 1984, 1989). The subjective belief that one can perform complex data analysis cannot substitute for demonstration of such, and those who readily argue that human introspection can be misleading must realize that they themselves are not exempt form this limitation.

5. *Description can inform prescription.*

Although it may be considered a philosophical sin to go from "is" to "ought," or from description to prescription, the boundary between the two in the history/philosophy of science covers only certain subdomains. What *has* succeeded and what *is* succeeding is predictive of what *will* succeed, and thus description has prescriptive implications. Were the past entirely nonpredictive of what one ought to do, it would follow that one could junk the scientific method altogether as a preferred procedure and consider all possible competitors (that do not abuse logic or rationality) equally viable. We do not, however, endorse a strong descriptive program. Specifically, we do not maintain that whatever scientists have done is normative (i.e., should be done), or that whatever has led to success should serve as a guideline. We certainly do not believe that all scientists have always functioned optimally and that all procedures all scientists have used are equally good or have the same prior odds of success.

6. *Small increments can produce big payoffs.*

Certain scientific efforts have little chance of success. For example, a scientist's effort to find a vaccine for a particular disease may be very unlikely to succeed. (We are not ignoring the point that a failed attempt may eliminate a faulty idea or that a community of researchers who pursue a research program may be likely to succeed eventually. We are simply pointing out that certain scientific efforts stand little chance of reaching a correct solution and that correct solutions may be slow in coming and require much effort.) Given the low probability of success for certain types of scientific efforts, small increments in the probability of success can make a big difference. For example, if the probability of success is .001, an increment of half a percentage point must be considered huge. It is not unreasonable to believe that better ways of combining information about the history of science can lead to guides or rules of thumb that enhance the probability of success in one or another type of scientific effort by at least a fraction of a percentage point, and that sometimes even small increments can amount to large differences. Thus, although, we anticipate fairly large absolute gains in some areas, one can hold more modest expectations and still believe that the type of science of science advocated here is worthwhile.

7. *Metatheory may be difficult but it is usually not impossible.*

Certain problems or research strategies one might pursue in the scientific study of science are not difficult either conceptually or methodologically, such as the proposal for actuarial grant evaluation we will outline below. Other problems, such as determining the manner in which science is growth-like, are clearly challenging but still may well be feasible.

Example: It follows from our arguments that one often needs an extensive and well organized data base on scientific episodes in order to test various hypotheses. One would like to have, for example, a data base that covers major episodes or events in the history of science over the course of decades or centuries. Additionally, one would like a sample that includes successful *and* unsuccessful scientific programs. One would like a directory of the players, which would include reliable descriptions of their stances on theories and related matters. Is the development of such a data base a difficult undertaking? An impossible one? Well, it is neither. It already exists! Frank Sulloway has compiled a data base that includes reliable ratings (as rendered by multiple expert historians) of all the major players and all the major episodes of scientific revolution over the last three centuries (the details can be found in Sulloway, 1990). Although Sulloway is clearly an exceptional person who has made an exceptional effort, his work shows that it is possible to accumulate the type of data base needed for certain of the forms of metatheoretical analysis we advocate.

As science is not only the best game in town, but also one of the biggest and most important, and as substantial resources are often expended on program evaluation, it is not outlandish to suggest that considerable effort and resources should be directed toward the development of new methods for evaluating science. From either a theoretical or practical viewpoint, the considerable effort needed to pursue the evaluation of science is hardly a fatal objection. One might contemplate the expense of poor or incomplete knowledge, and thus whether certain evaluative programs are likely to be worth the cost. For example, suppose the proposal that follows for actuarial grant evaluation, which could probably be developed and implemented at modest or low cost, could save 5% of currently wasted funds. Considering what is spent annually for scientific research, this would offset by manyfold the cost of development.

Some Illustrations

Representative Sampling of Scientific Events or Episodes

Many claims about scientific processes are inherently statistical (Meehl, 1984). For example, one or another method, approach, or strategy is said to characterize episodes of typical, successful, or preferred science. Laudan et al. (1986) list contrasting assumptions about scientific change that leading historians and philosophers have put forth. A remarkable number of these assumptions contain frequency claims. In the extreme case, of the 15 assumptions of hypotheses listed under a category for successor theories, every one of them contains either the words "seldom," "rarely," or "always." If one is hypothesizing a certain level of occurrence, one generally must obtain a representative sample to determine whether the claim is accurate (the exceptions obviously including such conjectures as never or always).

What should be sampled depends on the question asked, if one is formulating hypotheses about the growth-like nature of scientific theories, one will not care whether scientist Smith was the first or last born in the family, whether the royalty offered incentives for fixing the calendar, or whether Smith's relation with Doe helped to procure the grant. Determining the unit of measurement, establishing the parameters of the population, obtaining quasi-objective markers, and developing practical sampling methods are not trivial undertakings, but this does not lessen the *need* for representative sampling. If one wants to know how often something occurs across one or another domain of science, one must sample across that domain in a representative manner. There seems to be no way around this problem, and whatever the technical difficulties involved, they apply equally to those who obtain episodes haphazardly or who garner the data impressionistically. The philosopher who selects episodes to demonstrate or illustrate a point has not surmounted sampling problems and associated technicalities that impede one who wants to sample properly; he has ignored them.

Suppose one is interested in episodes in which most scientists converted to a new theory or set of ideas. Various assumptions have been made about such conversions. For example, there has been much debate over who converts, or who *first* converts. Is it predominantly the younger scientists? Those who have been less centrally involved in the development of the old theory? Those with extra-scientific beliefs in keeping with the newly proposed world order? There is also much debate over whether conversion to a new theory is slow or rapid, the extent to which critical experiments contribute to change, and whether selection is based primarily on track record versus potential. Argument and debate, founded on impression, have done little to settle these disputes, which are inherently statistical. Representative sampling would probably go a long way towards resolving such disagreements (Donovan et al., 1988; Hull, Tessner, & Diamond, 1978; Simonton, 1988).

Although there is dispute about the relative importance of track record versus potential in the selection of theories, there is general agreement that newly accepted theories may be less thoroughly tested than the theories they replace, an observation that raises deep and enduring questions about the nature of scientific progress and progression. First, to the extent that theories are selected on the basis of promise rather than performance, and to the extent that promise is not a perfectly discernible quality (which it obviously is not), there is the potential for error. The frequency of erroneous choice and, in particular, the mechanisms that do or could reduce its likelihood without overly compromising openness to new ideas are complex puzzles. Here we have a three-fold conjecture: (1) evaluations of promise are not as accurate as is typically assumed, (2) the frequency of good selection shows a strong negative correlation with the extent to which appraisals depend on promise versus track record, and (3) such judgments could be considerably improved through actuarial methods (see below).

In addition, there are longstanding debates about the manner in which new theories are growth-like. Need new theories have excess content? Solve anomalies that their predecessors could not while preserving the problem-solving capacities of their predecessors? Make predictions with greater precision? Increase conceptual clarity or depth of understanding? and so on. Again, we would contend that progress on such questions partly hinges on the quality of sampling.

Actuarial Methods for Aiding Scientific Appraisal

The following discussion will presume basic familiarity with the issue of actuarial versus clinical judgment (see Meehl, 1954; Dawes et al., 1989). The term "clinical" does not refer here to a clinical setting or practitioner, but rather to a method of judgment in which decisions are made in the head. In the actuarial method, decisions are based strictly on observed frequencies or empirical relations. The great majority of higher level evaluative work in science is conducted clinically. For example, those attempting to predict the success of proposed research (e.g., grant reviewers), determine the merit of completed research (e.g., journal editors), or evaluate theories, do not sit down with actuaries and ask them to get out their calculators. Rather, they take the available information and integrate it in their heads. We conjecture, given the limits in human cognition and the complexities of such evaluative work, that judgments of these types would be improved if conducted actuarially. We see movement towards actuarial evaluation of scientific products as among the early steps in the development of decision aids for higher level scientific judgments, as a starting point in the evolution of scientific method through which powerful aids for theory testing, evaluation, and possibly construction will emerge (see Faust, 1984).

We will discuss actuarial prediction of research productivity and of theories as two potential exemplars. In principle, it should be a simple matter to develop actuarial formulae for predicting the ultimate success of grant proposals (not whether they are funded, but whether they lead to productive research). For example, one could start with whatever dimensions raters currently use. One then examines the relations between these ratings and the outcome of research efforts, preserving and weighting (if necessary, which it may not be) the dimensions that turn out to be predictive.

One might ask what advantage there could be in this undertaking, as it may seem to reproduce what the grant reviewers are doing already. However, actuarial methods do not necessarily retain all of the initial variables or ratings, but only those that are predictive, and, if needed, can combine them optimally. Distinguishing the predictive and non-predictive variables and weighting the predictive variables optimally are among the tasks that give human judges such a difficult time. In a comparable study with radiologists on the determination of disease severity, Einhorn (1972) found that *some* of the physicians' ratings of particular dimensions were useful, but that their global or overall ratings were not predictive. Actuarial formulae based upon the radiologists' ratings, which selectively incorporated the valid ratings and combined them properly, did have some predictive power. For grants, ratings of various dimensions are usually already available, and the additional effort needed to compile such ratings and develop actuarial formulae (once methods for evaluating outcome have been developed) is almost trivial. In a comparable situation, one of the authors was able to develop actuarial formulae for less than $1,000.00.

As some grants are funded and some are not, one would look for instances in which the same grant was rejected at one foundation and accepted at another. One would thereby have access to outcome data despite highly variable decisions/evaluations, enhancing the chances of uncovering predictive relations. The critic who complains about fuzziness in measurement of outcome might consider that grant evaluators (who assumedly have developed their skills in part by observing relations between researchers' intended or specific actions and outcome) operate under the same flaws and limits. For example, if a certain outcome measure is useless, it will be equally useless to the actuarial method and the grant evaluator. Further, actuarial methods can be used to predict both objective and subjective measures of outcome. Studies could include objective and subjective outcome ratings, such as citation counts and independent experts' ratings of the merit of completed work. It would be particularly impressive if actuarial methods were more accurate predictors of both objective and subjective ratings, an outcome we consider likely.

We will next consider the actuarial evaluation of theories. This is a much more involved topic, and for further details one should see Faust (1984) and Meehl (1990a, 1990b, 1990c, [1992b]). We would alert the reader that those less well-versed in the philosophy of science may find the following material difficult.

In studying scientific theories the "scientific" metatheorist will investigate statistical correlations between their various properties and relations, both longitudinally and cross-sectionally. A rough classification of metatheoretical (theory-of-theorizing) predicates to be thus examined is:

I. Internal properties
    A. Formal
        1. Logical
        2. Mathematical
    B. Conceptual-Substantive
II. External predicates (relations)
    A. Empirical performance ("track record")
    B. Psychosocial

Space limitations forbid detailed expansion or defense of this outline, so we can only exemplify briefly. Each of these predicates would be subjected to a quantitative index, either "objective" or rated by judges. *Examples*: A *logical* property would be *interknitting*, to the extent there are multiple cross-connections among concepts by their appearing linked to other concepts via overlapping derivation chains (derivations of facts from postulates, where theoretical terms appear in several such sequences, "overlap"). A *mathematical* property is how detailed the theory is about signs of, and orderings between, derivatives of functions. A *conceptual-substantive* property is the type of theoretical entity postulated (substance? event? state? disposition? field? structure? particle? mentalistic?). An *empirical performance* property is the narrowness (= Popperian "riskiness") and accuracy (= observational closeness) of numerical predictions. A *psychosocial* property would be the rate of acceptance of a theory by highly prestigious scientists.

We do not expect, and certainly ought not demand, that a rigorous a priori demonstration be provided for a proposed numerical index of a property. While some sort of metatheoretical rationale (perhaps heavily dosed with quasi-consensual intuition) is desirable, the long term "justification" for index choice is the *empirical orderliness revealed*, as it is in other sciences. Smooth time curves and high inter-index correlations will tell us whether we are getting somewhere or not. Of course the first, big question to ask about a property index is how well it anticipates the long-term fate of theories: which ones become ensconced in textbooks and encyclopedias as "firmly established" (scientists often stop calling these clear winners "theories," labeling them as "facts") and which end up as "losers" (e.g., phlogiston, caloric, John B. Watson's theory of maze learning) mentioned in the textbooks, if at all, as historical curiosities.

To put a little flesh on these abstract bones, we take one "factual track record" property: narrowness and accuracy of numerical prediction. Meehl (1990a, 1990b, 1990c) has devised a crude index of corroboration in which the observational result of an experiment testing a theory's numerical prediction is put into a formula

$$C_i = (1 - I/S)\ (1 - D/S)$$

where I = the numerical interval tolerated by a theory, D = the deviation of observed value from the edge of the tolerated interval, and S = the *Spielraum* (the range of possible outcomes) antecedently available atheoretically. Suppose a casual conjecture predicts a correlation of .75 between two variables, with a theoretical tolerance of ±.15. We observe r = .67. The Spielraum is $.00 < r < 1.00$, if our background knowledge leads us to assume it will at least be positive. Then the corroboration index is $C_i = (1 - {}^{.15}\!/_{1.00})\,(1 - {}^{.08}\!/_{1.00}) = .78$. (Note that this has nothing to do with significance testing!) "Best" and "worst" case scenarios are used to standardize the crude index C_i, and the theory's mean value over experiments is taken as an index of its risky-accuracy track record.

The scientific fictionist (who rejects the reality of theoretical entities) would presumably be satisfied with finding strong predictive relations among these indexes since, for him, the theory is merely instrumental, a means to an end. "The purpose of theories is to predict and control facts," as the Psychology 1 cliché puts it. The scientific realist, on the other hand, is mainly interested in the *truth* of theories; the means–end relation is reversed, and the prediction and control of facts becomes the epistemic path to assessing the truth of theories. Since our position is realist rather than fictionist, we cannot avoid saying something more about this question.

Theories rarely being literally true (almost never, in social science), the leading concept of the realist is *verisimilitude* (Latin: "truth-likeness"). Some theories are nearer to the truth than others. Unfortunately the logicians have not as yet provided a rigorous explication of this concept that commands general assent, and some have doubts it can be done. Attempts to define verisimilitude through the relation between truth and falsity content (e.g., Niiniluoto, 1984,1987; Popper, 1962) have encountered various problems, and there seems to be general agreement that such approaches will not work (cf. references in Brink & Heidema, 1987; Goldstick & O'Neill, 1988). We conjecture that they have gone about it in the wrong way. There are several different (although related) respects in which a theory can err or be correct. For example, does it postulate the right kind of entities (particles, fluids, neurons, whatever)? If so, does it correctly state their causal or compositional relations (i.e., what is connected to what)? If it correctly asserts that one theoretical variable is functionally dependent on another does it get the direction of influence correct? If it does, is the function linear or decelerated? We can list an ordered set of such right-or-wrong features, Guttman scalable (or nearly so), and such a list has been offered by Meehl (1990c). A given theory, its postulates being explicitly formulated (if the scientist hasn't done this, then the metatheorist must), can be given a "verisimilitude score" by comparing it with the accepted theory of the textbooks, taken as a quasi-Gold Standard Criterion.

Logicians may be troubled by the possibility of an accepted theory being refuted later on. But this will not distress the actuarial metatheorist, who knows that psychologists routinely build and validate mental tests relying on fallible ratings, diagnoses, and developmental changes (the "bootstraps effect" in Cronbach and Meehl, 1955). We can ascertain which theory properties are statistically predictive of verisimilitude by correlating the property indexes with the verisimilitude index, despite the latter being itself imperfect. Some aspects of verisimilitude are more "intrinsically important" (to the realist) than others, but here we rely on the well known psychometric principle that for as many as 10 variables having positive manifold, two random weightings will yield highly correlated compositive indexes. For an extended discussion of verisimilitude, see Meehl (1990c, [1992a, 2002, 2004]).

Reviewers of this manuscript raised a number of the questions our proposal might generate. One reviewer cited Nickles (1986) to support doubts about the generalization of indexes, the notion being that evaluative criteria and preferable strategies are highly specific to area of inquiry. Concerns were also expressed about the stability of indexes over time, the vagueness of outcome criteria, and the suitability of the softer sciences, in particular psychology, to our approach. Some of these problems are general and are not compounded by the strategies we propose. For example, vagueness in outcome criteria is problematic across evaluative programs; formal indexes do not make this problem worse and instead may lessen it. Further, our proposal does not require the universal applicability of indexes or long-term stability, although were generalization and stability generally low the potential feasibility and power of our approach would probably be decreased (although, again, the same would apply to other evaluative strategies and methodological dictates). We anticipate all degrees of generalization and the need to adjust particulars across domains. For example, an index that reflects the conformity of outcomes to predictions in relation to riskiness should have broad application, but the standards one might set for theories are likely to differ considerably across, say, physics and psychology. In contrast, an index of reducibility downwards or upwards is likely to generalize less broadly. Such questions and possibilities bring us back to our overriding point: These types of science-of-science conjectures are exactly what we ought to be studying through more rigorous methods, rather than attempting to resolve them impressionistically and in the absence of decent data.

Addendum

Unfortunately, as it sometimes happens, Jones lost interest in his project and decided to pursue another line of work. It seems to us that psychologists, given their methodological sophistication, are among those best equipped to take up the project. It is our hope that someone will.

References

Arkes, H. (1981). Impediments to accurate clinical judgment and possible ways to minimize their impact. *Journal of Consulting and Clinical Psychology*, *49*, 323-330.

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

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. [Reprinted in P. E. Meehl, *A Paul Meehl Reader: Essays on the practice of scientific psychology* (pp. 9-30). (N. G. Waller, L. J. Yonce, W. M. Grove, D. Faust, & M. F. Lenzenweger, Eds.). Mahwah, NJ: Erlbaum, 2006.]

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.

Donovan, A., Laudan, L., & Laudan, R. (1988). *Scrutinizing science: Empirical studies of scientific change*. Boston: Kluwer Academic Publishers.

Einhorn, H. J. (1972). Expert measurement and mechanical combination. *Organizational Behavior and Human Performance*, *7*, 86-106.

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

Faust, D. (1986). Research on human judgment and its application to clinical practice. *Professional Psychology: Research and Practice*, *17*, 420-430.

Faust, D. (1989). Data integration in legal evaluations. Can clinicians deliver on their premises? *Behavioral Sciences & the Law*, *7*, 469-483.

Faust, D. (in preparation). The limits of scientific reasoning: Implications for the study and practice of science.

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

Hull, D. L., Tessner, P. D., & Diamond, A. M. (1978). Planck's principle: Do younger scientists accept new scientific ideas with greater alacrity than older scientists? *Science*, *202*, 717-723.

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

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

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.]
Available at http://www.tc.umn.edu/~pemeehl/

Meehl, P. E. (1973). Why I do not attend case conferences. In *Psychodiagnosis: Selected papers* (pp. 225-302). Minneapolis: University of Minnesota Press.
Available at http://www.tc.umn.edu/~pemeehl/

Meehl, P. E. (1984). Foreword. In D. Faust, *The limits of scientific reasoning*. Minneapolis: University of Minnesota Press. Available at http://www.tc.umn.edu/~pemeehl/

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

Meehl, P. E. (1990a). 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, N.J.: Lawrence Erlbaum Associates. [Reprinted in *A Paul Meehl Reader: Essays on the practice of scientific psychology* (pp. 445-486). (N. G. Waller, L. J. Yonce, W. M. Grove, D. Faust, & M. F. Lenzenweger, Eds.). Mahwah, NJ: Erlbaum, 2006.] Available at http://www.tc.umn.edu/~pemeehl/

Meehl, P. E. (1990b). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant using it. *Psychological Inquiry*, *7*, 108-141. Meehl's reply to the commentators. *Psychological Inquiry*, *7*, 173-180. [Reprinted in *A Paul Meehl Reader: Essays on the practice of scientific psychology* (pp. 91-167). (N. G. Waller, L. J. Yonce, W. M. Grove, D. Faust, & M. F. Lenzenweger, Eds.). Mahwah, NJ: Erlbaum, 2006.] Available at http://www.tc.umn.edu/~pemeehl/

Meehl, P. E. (1990c). *Corroboration and verisimilitude: Against Lakatos's "sheer leap of faith"* (Working Paper, MCPS-90-01). Minneapolis: University of Minnesota, Center for Philosophy of Science. Available at http://www.tc.umn.edu/~pemeehl/

Meehl, P. E. ([1992a]). Cliometric metatheory: The actuarial approach to empirical, history-based philosophy of science. *Psychological Reports, 71*, 339-467. [This is the first of three publications that Meehl originally intended to publish as a book. It was followed by: "Cliometric metatheory II: Criteria scientists use in theory appraisal and why it is rational to do so." *Psychological Reports*, 2002, *91*, 339-404; and, published posthumously, "Cliometric metatheory III: Peircean consensus, verisimilitude, and asymptotic method." *British Journal for the Philosophy of Science*, 2004, *55*, 615-643. —LJY] All three available at http://www.tc.umn.edu/~pemeehl/

Meehl, P. E. ([1992b]). The Miracle Argument for realism: An important lesson to be learned by generalizing from Carrier's counter-examples. *Studies in History and Philosophy of Science*, *23*, 267-282. Available at http://www.tc.umn.edu/~pemeehl/

Nickles, T. (1986). Remarks on the use of history as evidence. *Synthese*, *69*, 253-266.

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

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

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

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

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

Simonton, D. K. (1988). *Scientific genius*. New York: Cambridge University Press.

Sulloway, F. J. (1990, February). *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.

pdf by LJY, June 2012