2002

Using Meta-Scientific Studies to Clarify or Resolve Questions in the Philosophy and History of Science

David Faust
University of Rhode Island, faust@uri.edu

Paul E. Meehl

Citation/Publisher Attribution

Faust, D., & Meehl, P. E. (2002). Using Meta-Scientific Studies to Clarify or Resolve Questions in the Philosophy and History of Science, *Philosophy of Science*, *69*(S3), S185-S196. doi: 10.1086/341845
Available at: https://doi.org/10.1086/341845

Using Meta-Scientific Studies to Clarify or Resolve Questions in the Philosophy and History of Science

David Faust†
University of Rhode Island

Paul E. Meehl
University of Minnesota

More powerful methods for studying and integrating the historical track record of scientific episodes and scientific judgment, or what Faust and Meehl describe as a program of meta-science and meta-scientific studies, can supplement and extend more commonly used case study methods. We describe the basic premises of meta-science, overview methodological considerations, and provide examples of meta-scientific studies. Meta-science can help to clarify or resolve long-standing questions in the history and philosophy of science and provide practical help to the working scientist.

1. Introduction. As a graduate student in psychology about 25 years ago, one of the authors (DF) was sent by his major professor to the head of the philosophy department to discuss certain technical issues. At that time, this author presented the basic features of a still somewhat tentative proposal for meta-science, which the other author (PEM) has co-developed and now refers to as the "Faust-Meehl Thesis." As we will describe, the Faust-Meehl Thesis involves a theoretical rationale for, and the design of, more rigorous methods for studying scientific episodes in order to assist in the understanding and integration of the massive historical track record. The program has both descriptive and prescriptive aims. Upon hearing

†Send requests for reprints to the authors. David Faust: Department of Psychology, 10 Chaffe Rd., Suite 8, Kingston, RI 02881. Paul Meehl: Department of Psychology, 75 E. River Rd., Elliot Hall, Minneapolis, MN 55455.

Philosophy of Science, 69 (September 2002) pp. S185–S196. 0031-8248/2002/69supp-0017$10.00

the proposal, this philosopher simply stated, "If you are correct, then my life work has been a waste and I am out of business."

It was, and remains, the conviction of both authors that this pessimistic pronouncement was wrong on both scores and that the meta-scientific approach or program that we will describe should have just the opposite effect, that is, that it will sharpen traditional problems and create new ones involving issues that are often central to historians and philosophers of science, leading to many productive undertakings. These problems and questions involve such matters as: What features of theories predict their long-term survival? To what extent are these features similar across disciplines and domains? Stated differently, meta-science should provide rich and hardy grist for the mill of historians, logicians, and philosophers of science.

In the article that follows we will discuss the potential benefits of applying more rigorous methods to the analysis of the historical track record, present certain basic premises of our meta-science program, discuss its rationale and aims, and present some examples of potential applications. Space limitations necessitate a dense presentation that might sometimes seem inadequately attentive to methodological obstacles and objections; various sources provide more detailed descriptions of the premises, aims, and potential methods of meta-science, as well as our thoughts about certain objections and practical problems (Faust 1984; Faust and Meehl 1992; Meehl 1983, 1992a, 1992b, 1999).

2. Methodology for the Study of Science. The major current approach to the study of science is the case method, which has yielded many insights and is seemingly irreplaceable for certain purposes. However, there are two fundamental reasons why this approach may not be the method of choice for certain types of problems or questions, at least when used predominantly or in isolation.

First, the data base of scientific episodes or occurrences is *massive* and growing rapidly. Science is BIG, and it is nearly impossible for anyone using the case study method to master and continuously track more than a relatively small proportion of this data base.

Second, relations between the methods that scientists employ and the outcome of their efforts are largely probabilistic, not deterministic. Much of the methodology that scientists use is not, strictly speaking, rule bound, but more so follows from rules of thumb, principles, or guides, many of which can lead to inconsistent or even opposing actions (e.g., start by simplifying versus start holistically). Good or even excellent methods do not guarantee success, nor do bad or poor methods always lead to failure. One might crudely classify methodology and outcome into a two-by-two table, with one dimension representing method (good versus bad) and the

other representing outcome (good versus bad). It is evident, given the massive data base of scientific episodes and because the relations between method and outcome are inherently probabilistic or statistical, that we could fill all four cells of the table with many entries, even if good method was much more likely to lead to a positive outcome than poor method. Consequently, for nearly any descriptive or normative program, no matter how sound, the proponent can find many supportive instances (although one might have to search much harder and more selectively in the case of some of these programs than others).

Additionally, the same scientific procedure or methodology can produce inconsistent or varying levels of success. The relationship here is the one to the many. Also, different procedures can lead to the same outcome, the relation here being the many to the one. Again, this speaks to the statistical nature of the relations between scientific procedures and outcome.

Consider also the features of theories that are deemed desirable. Among the lists of such features that are commonly put forth, there is much, but certainly not complete, overlap. There is certainly not agreement about which features to assign the greatest importance or weight, or which should countervail one or more of the other features when they are inconsistent or different features favor competing theories.

Take the following abbreviated list of desirable features of theories. The list might include parsimony, which itself can be divided into a number of characteristics, such as simplicity of explanation or the fewest postulates per observation statement. The list might also include novelty in relation to numerical precision, that is, some variation of Popperian risk or Salmonian "damn strange coincidence." To these we could add rigor, qualitative diversity or breadth, reducibility upward or downward, and elegance or mathematical beauty.

No credible philosopher of science has claimed that any one of these features is a sure-fire guarantee of truth, or even a high level of verisimilitude. Nor has any credible philosopher claimed, despite a strong emphasis on one or two features, that any one always trumps over all the others. Thus, anyone who relies on any one of these features to appraise a theory's status must be claiming statistical relations between the presence or standing on that feature and the success of the theory or its verisimilitude.

The only essentially unambiguous case is the trivially simple one in which Theory A beats Theory B on all features. Commonly, however, the features themselves are inconsistent within and across theories, creating a potential judgmental dilemma. For example, Theory A may have excellent parsimony but modest rigor; or Theory A may surpass Theory B on some features, but for other features the opposite might hold.

Again, given the massiveness and probabilistic nature of the historical

track record, it is possible to identify many positive or negative instances for nearly any set of preferences proposed. In this context, case study becomes a method for refuting extreme claims of the type that almost nobody makes. For example, in *Realism and the Aim of Science* (1983), Popper cites multiple examples of theories that were abandoned quickly due to clear falsifiers. What does this refute? Has anyone claimed something like: "No scientific theory has ever been quickly abandoned because of what appeared to be clear falsifiers"?

If the claim instead is that scientific episodes should conform to certain characteristics, or that a certain approach will often or tend to yield a certain outcome, then selective illustrations are not helpful and different methods are needed. Given the size and heterogeneity of the historical data base, it is possible to pile up examples for nearly any program, even if the description is far from typical or the normative suggestions are less than optimal, if not relatively poor. If there are tens of thousands of episodes from which to collect examples, then even an approach that occurs or works 1% of the time will lead to hundreds of conforming instances.

Most importantly, methods for studying the historical track record need to incorporate some form of representative sampling of scientific episodes. Obtaining representativeness will generally require random sampling of a sufficient number of episodes (although this number may not need to be nearly as large as one might suppose). If we want to know what and how often something occurs, representative sampling is often, far and away, the most powerful method.

Many claims about science contain frequency statements or assertions that are fundamentally statistical. It is informative, for example, to review Laudan et al.'s (1986) list of contrasting assumptions about scientific change. Of the 15 assumptions or hypotheses listed under the category for successor theories, *every one* of them contains such terminology as "seldom," "randomly," or "always."

Why not just collect and combine episodes in the history of science on the basis of trained, expert judgment? First, although not literally true, the quality of conclusions are constrained by the quality of the data upon which they are based. Absent representative sampling, one lacks the data base needed to best answer or resolve these types of inherently statistical questions. The typical case study method does not capitalize on the far more powerful methodology that is available for obtaining representative samples and is unlikely to produce the needed representativeness. Further, in some instances, the case study method is directed toward identifying or accruing instances that illustrate or support a position, and therefore is likely to produce skewed, or grossly skewed, samples.

Second, optimal or improved integration of large and complex data bases is likely to be facilitated by decision aids that supplement the power

of the unassisted human mind. As legion research shows (e.g., see Faust 1984), the capacity of the unaided mind is greatly strained, if not far overburdened, when asked to optimally combine multiple variables with probabilistic relations to outcomes. The unaided human mind simply does not perform these types of operations or computations well. As Meehl (1986, 372) has stated:

> Surely we all know that the human mind is poor at weighting and computing. When you check out at the 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.

Although it might be argued that the case study method will usually be effective in identifying major differences in rate of success, matters become far more difficult when one wants to know just how often an approach succeeds across applications; or if one method beats another by a margin of, say, 25%, 10%, or 5%; or if one approach works somewhat better than another in some situations but not in others. The problem of subjective discernment can become especially difficult because, among other things, the less successful method may have been used far more often than the more successful method, leading to an absolute number (versus proportion) of positive outcomes that exceeds that of the more effective approach. Even relatively small differences in success rates can be of great importance to working scientists, especially when these probabilities are joined across scientific undertakings. For example, when the probabilities are multiplicative, five attempts with a 5% versus a 2% rate of success has a many-fold greater chance of achieving a positive outcome.

The problem of integrating episodes in the history of science and determining probabilistic associations between procedure or theory features and long-term outcome is worse than this, however, because one may well have to assign weights to the variables and also examine inter-relations or configural patterns among the variables. For example, although success with novel prediction may generally be a more powerful indicator of a theory's fate than parsimony, this may not hold true when the range of phenomena for which accurate prediction is achieved is very narrow and the alternative theory shows not only greater parsimony but also much greater breadth; alternatively, the relative weight that should be assigned to one or another variable may depend on the standing of other variables, that is, it may depend on patterns or configural relationships. To give what might be an overly simplified example for purposes of clarity, parsimony might count for nothing if novel prediction is nil, might count more if a theory also shows good rigor, and perhaps should be weighted heavily if the theory shows good standing on breadth. A quote from Dawes, Faust, and Meehl (1989), in follow up to Meehl's statement quoted above, illus-

trates the difficulties encountered when attempting to perform these types of mental operations subjectively:

> It might be objected that this analogy, offered not probatively but pedagogically, presupposes an additive model that a proponent of [subjectively accomplished] 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 weight a complex composite of these variables. (1672)

We, of course, do not mean to compare the evaluation of theories to supermarket pricing. Our example is intended to illustrate the difficulties encountered when one attempts to subjectively integrate multiple variables with probabilistic relations to outcome, variables which may act differently when combined and weighted in different ways or in different configurations. Thus, in addition to representative sampling, methodology designed to assist in the analysis and integration of such data bases (e.g., statistical methods such as multiple regression) can greatly bolster our judgmental accuracy and understanding.

3. Description and Prescription. More powerful methods for studying and integrating the historical track record can help clarify or resolve longstanding questions in the history and philosophy of science and provide practical help to the working scientist. Perhaps the most fundamental reason why better description helps the practicing scientist is that what was or is most successful in the past has value for predicting what will succeed in the future. If the past were entirely non-predictive on these matters, we could junk the scientific method completely. Imagine if we believed that a statement like the following were justified, "Just because control groups have helped us in thousands of past experiments, and just because this situation closely resembles the types of problems for which control groups have worked before, there is no basis to assume that a control group will help in this instance." Or, more broadly, "The past usefulness of control groups for decades and across thousands of studies and broad domains does not allow us to predict that control groups will assist us in future studies." Scientists, of course, consider the past track record of methods and approaches all of the time when planning or conducting new work. However, greater precision and accuracy, especially around matters that require complex data integration (e.g., which factors in which combination

best predict the long-term fate of theories) should provide improved guidance.

4. Two Illustrations. A variety of problems might capture the attention of the meta-scientist, especially problems in the history and philosophy of science that require the integration of complex data. For example, representative sampling and statistical analysis might be applied to the study of scientific change, or to the association between scientist's methodological preferences and the success of their efforts. Given space limitations, we will limit ourselves to a discussion of two possible areas of study.

4.1. Grant Evaluation. Grant evaluation involves prediction under conditions of uncertainty, that is, reviewers attempt to predict the outcome or utility of proposed, but yet to be conducted, research studies or programs. Presently, grant evaluation is almost always conducted through some form of data integration that rests substantially or mainly on subjective judgment. This is the case even should these evaluations involve assigning ratings to various dimensions and then adding up scores on these dimensions or using another means for formulating some type of global ratings, because the selection of the dimensions and the scheme for combining the dimensions are, themselves, subjectively derived. How would the meta-scientist proceed in this domain?

One might initially identify a range of variables that seem relevant in judging the quality of grant proposals or in predicting their success. It would be sensible to start this process by eliciting the beliefs and impressions of qualified scientists, particularly those considered expert in grant evaluation. We would begin by generating a list of evaluative features that, if anything, is overly inclusive. Mistakenly including variables on the candidate list should not be too serious an error, because proper analysis will help us identify those that do not work or are unnecessary (i.e., that are non-predictors, weak predictors, or redundant predictors). In contrast, the failure to include potential predictors may represent missed opportunities.

The various grant proposals are rated along these dimensions, taking steps to ensure that the ratings are reliable or consistent across evaluators. Classical psychometrics provides formulas for such questions as how many judges must be pooled to achieve a desired level of reliability, the constraints that level of reliability sets on validity, and the like. We then examine, through the proper mathematical procedures (e.g., multiple regression), the relations between standing on these background variables and outcome, that is, the fate of the executed research project. At this stage, we will probably prefer to work with archival data. With archival data, we need not await outcome, can examine a long enough time period after completion of the research to make more accurate and trustworthy

judgments of success, and can avoid cases with more ambiguous outcomes, or for which success is particularly difficult to rate.

The mathematical analyses will tell us which variables are and are not associated with outcome, how strongly they are associated, and what variables in what combination or weighting scheme maximize predictive accuracy. For example, it may turn out that a researcher's past success is a far more powerful predictor than institutional affiliation or the thoroughness of the literature review. We might find that a substantial number of variables all contribute independently to prediction (an outcome that, for technical reasons we cannot enter into here, we consider unlikely); that many of the variables are redundant and that only a relatively small subset are needed to maximize prediction; and that some variables generally believed to be good predictors are not and that other variables often considered to be of secondary importance are among the best predictors. It might be that the useful variables can simply be added up and weighted similarly to maximize predictive accuracy, that differential weighting is needed, or that combinations, or complex combinations of these variables must be utilized. Of course, we do not know what we might find—we may just end up "confirming" what was assumed all along—but this is the point of doing such studies. Of interest, a large body of research shows the feasibility of conducting these types of analyses of human or expert judgment, although this work has not yet been applied to the study of higher level scientific judgments. Further, this research on judgmental processes often reveals substantive discrepancies between subjective appraisal, weighting, and integration of variables in comparison to what statistical and mathematical analyses show is optimal (Faust 1984; Meehl 1954; Grove and Meehl 1996).

Further analyses could be conducted to determine whether the originally derived predictors or predictive formulae are stable and generalize to new cases within the same domains, and the extent to which they can be applied to other domains. For example, the variables that predict outcome in a novel area of psychology may well differ, or differ markedly, from predictors in an advanced area of theoretical physics.

A critic may raise various objections to this proposal. For example, doesn't such an approach, which starts with dimensions identified by raters and their ratings of these dimensions, duplicate what is already done? The answer is that it might, but it might not. Grant evaluation involves more than identifying relevant dimensions and rating them, it also involves integration of the ratings. Further, the dimensions selected for evaluation may or may not be predictive. Formal approaches can help to determine whether the variables that subjective appraisal leads us to value are good predictors, whether other predictors that are considered weaker or inferior might have greater value than believed, the extent to which predictors are

redundant and therefore add little or nothing to predictive accuracy, and how to best combine these variables. The end product may match, or greatly differ from, what we believe or what we are doing subjectively. Many related studies of expert judgment show that these statistical or mathematical methods almost never lead to inferior overall prediction in comparison to subjective data integration and, instead, often bolster predictive accuracy, sometimes substantially (Dawes, Faust, and Meehl 1989; Grove and Meehl 1996).

A telling example is provided by Einhorn's (1972) study. Einhorn had radiologists rate biopsy slides along a series of dimensions that they believed were indicators of disease severity. The radiologists also provided a global rating of disease severity. In the case of the medical condition under consideration, severity should be related to survival time. Sadly, outcome data were ultimately available because the patients had terminal illness. The radiologists' global ratings of disease severity showed no relation with survival time post-biopsy. Ironically, however, statistical analysis indicated that some of the variables that they rated, although not all, were associated with survival time. Further, a statistical combination of this subset of valid predictors did achieve modest accuracy. Einhorn's study showed that the pathologists were able to generate useful data, but that they themselves did not make proper use of these data. The failure of their global ratings to predict outcome and the contrast with statistical methods suggest that the radiologists had difficulty discerning which of their own ratings, or the dimensions that they rated, were predictive, as well as determining how to best combine the information. Einhorn's seminal findings have gained considerable support across a range of decision domains (see Connolly, Arkes, and Hammond 2000; Faust 1984).

It is certainly possible that grant evaluators have similar difficulty distinguishing the predictive from the non-predictive dimensions and combining the information optimally. Research of this type also permits study of new or novel predictors that might not normally be used. Additionally, as knowledge expands, new variables or combinations of variables can be uncovered that are otherwise difficult to conceive of or anticipate and that are problematic to evaluate subjectively. For example, an index of a researcher's past success might be incorporated into an overall predictive formula and might include cumulative ratings of such variables as consistency of work quality, citation patterns, and upward or downward trends over time.

Another objection might be that such research is seemingly limited to grant proposals that are funded, which reduces variation in ratings and perhaps outcome, both of which can hinder the effort to uncover predictive variables. For example, if we are more or less limited to grants that have been assigned fairly uniform, positive ratings, how can we determine

how variation in rating (of which there is little) is related to outcome? Further, if grant evaluators generally do make good judgments and give lower rating to proposals that usually would fail, study of the proposals that are funded might not reveal these proper judgmental practices. However, if we do not limit analysis to a single agency we can likely overcome this problem: many grants that are rejected by one agency are accepted, unchanged, by another agency. We are not suggesting that research on this and other meta-scientific topics is always or necessarily easy, only that it is often feasible. Given the size and importance of science in current society, time and resources dedicated to studies that can improve the effectiveness of scientific endeavors is likely to be a wise investment.

Concerns might also be raised about methods used to rate the outcome of the funded research. One approach is to obtain both objective and subjective ratings of outcome. Objective ratings, for example, could include citation counts. Subjective ratings might include the evaluations of experts. Superiority, and especially clear superiority, across both objective and subjective ratings would create a potentially strong basis for declaring a winner (i.e., the original ratings of the grant evaluators versus the statistical predictions). To the extent that the appraisal of outcome is fuzzy, this is not necessarily an argument for or against meta-scientific methods in comparison to current methods of grant evaluation. That fuzziness equally confronts those using present methods, and hence is a very questionable basis to argue for current methods over meta-scientific methods.

4.2. Evaluation of Theories. The aim here is to develop predictors of the success of theories or their long-term fate. One might again start with a list of properties or indices and then, via study of the historical track record, analyze relations between standing on these variables and theory success. For example, an index might be designed to evaluate, roughly speaking, Popperian risk, or predictive accuracy in relation to risk (for further details on this and other possible indices, particularly in regard to methodological issues and potential objections, see Meehl 1997, especially 415–417). The index might include, first, range of possible (or plausible) outcomes. To illustrate, the typical experiment in psychology has two or a few possible outcomes (e.g., Variable A will or will not be related to Variable B), whereas earlier studies in chemistry which involved predicting the number of molecules in a mole had an enormous range of possible outcomes. The index would also include the match or discrepancy between predicted and obtained outcome, that is, the closeness of fit. For example, an outcome that is relatively close provides much stronger support for a theory with a very large versus a much smaller range of possible outcomes. Thus, one examines the discrepancy between predicted and ob-

tained outcome in relation to range of possible outcomes or risk. One places range of possible outcome in the denominator and discrepancy between predicted and obtained outcome in the numerator. Consequently, the greater the range of possible outcomes and the smaller the difference between the prediction and outcome, the smaller the obtained number. The result can be subtracted from 1 for ease of interpretation, so that the higher the number, or the closer it is to 1, the better the outcome. The index can be calculated across relevant studies and a cumulative rating derived. Other indices might rate such dimensions as qualitative diversity and parsimony.

Working with a range of potential variables or indices of theory status, some traditionally described and others perhaps less traditional or not yet developed, one can examine their predictive power, how variables are best combined, and how to manage inconsistencies among predictors. For example, using various indices in varying combinations, the performance of competing theories could be plotted over time. It would be of interest to determine whether examination of performance curves or separations between theories might allow a winner to be identified, and how the accuracy and timeliness of such judgments compared to that of the scientific community. For example, in some instances these meta-scientific indicators of theory status might identify winners or loses sooner, or much sooner, than other means. Again, much of this initial research would likely capitalize on archival data, for example, clear cases in which one theory wins out over another or achieves long-term success.

A critic might object here that even should it be possible to identify characteristics associated with the long-term fate of theories in a particular domain or subdomain, any such indicators are unlikely to generalize to other scientific domains. For example, variables that predict the long-term fate of theories in a particular branch of biology might be useless for predictions within astronomy. We would anticipate a relatively high level of generalizability for some variables (e.g., predictive accuracy in relation to risk) and less, or considerably less generalization for other indicators; but, more so, we believe that such matters are difficult to anticipate, which is exactly why such studies are needed. Potential worries about the availability of the needed historical track record have been nearly resolved by Sulloway (1996) who, almost single-handedly, has demonstrated the feasibility of generating the data bases required to perform meta-scientific studies. Also, to facilitate and simplify initial efforts, the meta-scientist could work in more delineated domains by studying mini-theories. Mini-theories in restricted domains can number in the hundreds (e.g., digestion) or thousands (e.g., human genetic mutation). Philosophers may be mistaken in focusing so heavily on the "grand theories" (e.g., Kepler, Darwin, Einstein). The supply of mini-theories is plentiful.

5. Conclusion. The capacity to think about thought was a major step forward in human intellectual development. Significant advance is often signaled or achieved when what has been the highest level of thought becomes the subject matter upon which intellectual operations occur. As data are the subject matter for theories, theories and other scientific products are the subject matter for meta-theory and meta-science, organized and directed by methods that, in large part, remain to be developed. However, we believe that the era of meta-science is not far off and that it will make significant, if not revolutionary, contributions to the history and philosophy of science, and to the work of the practicing scientist.

REFERENCES

Connolly, Terry, Hal R. Arkes, and Kenneth R. Hammond (eds.) (2000), *Judgment and Decision Making,* 2nd ed. New York: Cambridge University Press.

Dawes, Robyn M., David Faust, and Paul E Meehl (1989), "Clinical Versus Actuarial Judgment", *Science* 243: 1668–1674.

Einhorn, Hillel J. (1972), "Expert Measurement and Mechanical Combination", *Organizational Behavior and Human Performance* 7: 86–106.

Faust, David (1984), *The Limits of Scientific Reasoning.* Minneapolis: University of Minnesota Press.

Faust, David and Paul E. Meehl (1992), "Using Scientific Methods to Resolve Enduring Questions within the History and Philosophy of Science: Some Illustrations", *Behavior Therapy* 23: 195–211.

Grove, William M. and Paul E. Meehl (1996), "Comparative Efficiency of Informal (Subjective, Impressionistic) and Formal (Mechanical, Algorithmic) Prediction Procedures: The Clinical-Statistical Controversy", *Psychology, Public Policy, and Law* 2: 1–31.

Laudan, Larry, Arthur Donovan, Rachel Laudan, Peter Barker, Harold Brown, Jarrett Leplin, Paul Thagard, and Steve Wykstra (1986), "Scientific Change: Philosophical Models and Historical Research", *Synthese* 69: 141–223.

Meehl, Paul E. (1954), *Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence.* Minneapolis, MN: University of Minnesota Press. Reprinted with new preface, 1996, by Jason Aronson, Northvale, NJ.

——— (1983), "Subjectivity in Psychoanalytic Inference: The Nagging Persistence of Wilhelm Fleiss's Achensee Question", in John Earman (ed.), *Minnesota Studies in the Philosophy of Science: Vol. 10, Testing Scientific Theories* Minneapolis: University of Minnesota Press, 349–411.

——— (1986), "Causes and Effects of My Disturbing Little Book", *Journal of Personality Assessment* 50: 370–375.

——— (1992a), "Cliometric Metatheory: The Actuarial Approach to Empirical, History-Based Philosophy of Science", *Psychological Reports* 71: 339–467.

——— (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.

——— (1997), "The Problem is Epistemology, Not Statistics: Replace Significance Tests by Confidence Intervals and Quantify Accuracy of Risky Numerical Predictions", in Lisa L. Harlow, Stanley A. Mulaik, and James H. Steiger (eds.), *What If there Were No Significance Tests?* Mahwah, N.J.: Erlbaum, 393–425.

——— (1999), "How to Weight Scientists' Probabilities is not a Big Problem: Comment on Barnes", *British Journal for the Philosophy of Science* 50: 283–295.

Popper, Karl R. (1983), *Postscript (Volume I): Realism and the Aim of Science.* Totowa, N.J.: Rowman & Littlefield.

Sulloway, Frank J. (1996), *Born to Rebel: Birth Order, Family Dynamics, and Creative Lives.* New York: Pantheon Books.