



Commentary

The philosophical legacy of Meehl (1978): confirmation theory, theory quality, and scientific epistemology

J. D. Trout

Philosophy Department and the Parmlly Hearing Institute, Loyola University Chicago, 6525 North Sheridan Road, Chicago, IL 60626, USA

Abstract

The reach and impact of Paul Meehl's work is extraordinary. Focusing on his "Theoretical Risks and Tabular Asterisks", I trace three consequences of his findings. The first is the influence of his work on confirmation theory in the philosophy of science, in which he provided a more sophisticated alternative to Popperianism, despite some affinities with it. The second is a clear focus on the evaluation of theory quality as an explanation for the success of hard vs. soft theories. The third is a very deep critique of the practices of contemporary epistemology; his research recommends the replacement of demonstrably unreliable, subjective judgments about justification and knowledge with simple predictive models that outperform human experts. This is an impressive array of intellectual contributions to psychology and philosophy. As influential as his work was, it is just now beginning to receive attention commensurate with its merits.

© 2004 Elsevier Ltd. All rights reserved.

Keywords: Confirmation theory; Theoretical Risks and Tabular Asterisks; Severe testing; Hard vs. soft theories

Philosophers are all too familiar with the rise and fall of intellectual fashion. Without decisive standards that might have signaled a theory's success or failure, new methods, principles and theories would be greeted with excitement, only to die from an illness never diagnosed, followed by a scholarly post-mortem deemed too dull to sustain. But philosophers of science in the 20th century found reason for optimism, and the promise of emulation, in the bracing methodological disputes in psychology, where advances seemed to give traction to statistical approaches to theory testing. So, it was disconcerting when we heard from Paul Meehl, a philosopher and psychologist at the top of his field, that the choice of statistical method might itself be an artifact of fashion, and more importantly, one that impeded progress in a proper science of the mind.

As Meehl put the point:

It is simply a sad fact that in soft psychology theories rise and decline, come and go, more as a function of baffled boredom than anything else; and the enterprise shows a disturbing absence of that *cumulative* character that is so impressive in disciplines like astronomy, molecular biology, and genetics. (1978, p. 807)

Contributing to this unprincipled, directionless drift, Meehl thought, was the rampant use of significance testing in the "soft" areas of psychology: clinical, school, community, social, counseling, and personality. One might quibble about the appropriateness of occupants on this list; the important point is that there is a potentially positive lesson here. If, as Meehl claims, significance testing is a "poor way of doing science" (1978, p. 806), a better method might advance these areas of psychology. Indeed, there would seem to be much room for improvement, as these soft areas of psychology "are scientifically unimpressive and technologically worthless" (p. 806).

1. Confirmation theory

Among philosophers, "Theoretical Risks and Tabular Asterisks" is probably best known as a trenchant attack on significance testing. Philosophers knew of it because it developed themes that are historically located in the philosophy of science—in particular, in confirmation theory. At the time of its publication, confirmation theory was full of proposals about the conditions for a good test. But it was unclear how any of them related to the actual contexts of theory development.

E-mail address: jtrout@luc.edu (J.D. Trout).

“Theoretical Risks and Tabular Asterisks” was much more than an exercise in confirmation theory, because it documented the ways in which this purportedly poor methodology retarded scientific progress. It also documented how the quality of a theory was, in practical terms, inseparable from the quality of a test it could offer. Meehl’s hope was to bring a quantitative rigor to theorizing and experimentation about the mind—not for the sake of formalism, but for the discipline it imposes on the flabby temptations of interpretation.

Meehl’s chief complaint, though he had many others, was that significance tests have unacceptably low severity. Low test severity exists when the probability is high that the hypothesis passes the test even when the hypothesis is false. The concept of a severe test dates at least to Karl Popper, one of the two figures memorialized in the title of Meehl’s landmark paper. Popper claimed, famously, that induction was not a rational procedure for theory choice. According to this view, theories are never confirmed by observations, and inductive arguments on behalf of theories are never justified. The reason concerns the high premium Popper placed on certainty. With the potentially infinite number of theories that could similarly predict the data, you can never be certain which theory is actually responsible for the positive outcome. It is important to recognize that for Popper, the refutation of a theory, and so its falsification, was the sole criterion of rational theory choice. As a result, nothing could be inferred from the corroboration of a theory, its survival under attempts to falsify it. But a refutation—a negative result of an experimental test—decisively disproves the theory. Though Popper introduced the notion of corroboration to capture the intuition that a positive outcome is not completely irrelevant to the test of a theory, it did little to clarify the incremental support that accumulates in valued areas of science.

This account of theory choice left an important tension unresolved. Popper’s criteria of theory choice entailed that we have no more reason to believe a theory that has survived many tests and one that has not been tested yet. So, Popper’s account of corroboration leaves us unable to decide between a tested but hitherto unfalsified theory and one that has not been tested at all, despite the fact that the former preference seems to offer the superior strategy.

In addition to these niceties of Popper scholarship concerning corroboration, there was a looming problem of the holistic character of theory testing: When a theory failed to pass a test, its failure might have nothing to do with the quality of the theory itself, but rather with the auxiliary hypotheses used in testing it. No hypothesis gets tested in isolation. This fact rendered ineffectual Popper’s effort to use the “falsifiability criterion” as an instrument to identify and derogue pseudo-scientific views.

Despite his reliance on Popper in “Theoretical Risks and Tabular Asterisks”, Meehl’s need for the Popperian architecture does not go deep. Meehl’s attachment to Popperian falsificationism was opportunistic, prompted by the desire to distinguish between scientifically respectable tests

and disreputable theories. By contrast, Popper’s aim was much more sweeping—to distinguish between science and pseudo-science. Meehl’s attraction to falsificationism is easily replaceable with the ordinary standards of induction that honors the important role of probability in science. After all, Meehl’s complaint about soft psychology was not that its theories could not be empirically confirmed, but that the intrinsic features of the subject matter, along with the weakness of the methodology, made the problems of theory choice very formidable using the standard statistical tools. But not insurmountable. In the end, the same might have been said of the scientific tools used to test evolutionary theory, and yet evolutionary theory has overcome those obstacles. (Indeed, it may be symptomatic of Meehl’s departure from falsificationism that he conceded the scientific status of evolutionary theory, while Popper did not.)

So, a retrospective appreciation of Meehl’s arguments in “Theoretical Risks and Tabular Asterisks” frees Meehl of an untenable use of falsificationism. At the same time, Popper and Meehl shared the quest for standards of a severe test. This retrospective reading of Meehl’s classic allows us to explain why Meehl focuses on soft theories rather than ones that are clearly pseudo-scientific. His aim is to improve methodology and theorizing, and it is soft theories that are distinguished by their practical resistance to severe test. (For a wonderfully clear introduction to, and criticism of, Popper’s falsificationism, see [Godfrey-Smith, 2003](#).)

2. Progress in hard and soft psychology

Ultimately, Meehl approved significance testing in some cases—in cases where all you are looking for is a difference. The cases in which it worked best were atheoretical: Was the antibiotic effective? It remains an interesting question why significance testing is used with such apparent success in the developed areas of perceptual and cognitive psychology. I take up this question in [Trout \(1998, 1999\)](#), and conclude that the answer is already implicit in [Meehl \(1978\)](#). In particular, Meehl’s silence on the use of significance testing in perception and cognition signals his recognition that its application is not as problematic as in the “softer” areas of psychology. The difficulties are, therefore, not intrinsic to significance testing, but depend on the conditions of its application—the quality of the theories that significance testing it is applied to.

In order to illustrate this dependence of successful theory-testing upon theory quality, Meehl pressed the following uncomfortable observation: When you do not have a good theory, a significance test easily, and illicitly, turns into a search procedure. And when you have a good theory, you do not need significance testing; a host of other methods can be used. Meehl commonly points out that the most successful sciences—physics, chemistry, and biology—do not use significance testing. For our purposes, however, the important question is, could they? It would be interesting to

know, with greater certainty, how Meehl would explain the success of significance testing in perception and cognition, but “Theoretical Risks and Tabular Asterisks” leaves little room for doubt—it would be explained in terms of theory quality.

3. Epistemology and intellectual discipline in quantitative hands

Meehl often characterized himself, in “Theoretical Risks and Tabular Asterisks” and elsewhere, as a “dustbowl empiricist”. His philosophical views gave epistemological priority to observation. This is understandable, because his home science was psychology, and much of “Theoretical Risks and Tabular Asterisks” set out reasons to distrust many psychological theories. So, observational outcomes really counted, for Meehl. This focus led to Meehl’s important work on actuarial versus clinical judgment. When you have theories of unproven success and a science without technological value, it is wise to stick to less problematic sources of information, such as observational outcomes, rather than the dramatic speculation of grand theories. When the observational outcomes are correlated with demographic (or other) categories, a powerful linear model can be constructed to outperform human judgment on a vast range of tasks. Paul Meehl was the first to publicize this fact and underline its importance.

Until recently, however, this groundbreaking work had no impact on epistemology. This neglect has been damaging, but interest in these lessons is growing; there is much to be learned by philosophy from that area of psychology that proposes to improve human reasoning (Bishop and Trout, *in press*). For example, research in cognitive psychology on judgment under uncertainty, fast and frugal heuristics, and linear predictive modeling, has produced a wealth of findings about how to improve human reasoning. The general lessons are that unaided judgment can be treacherous, but there are simple rules or models that routinely outperform human experts on a wide range of important tasks: college and medical school admissions, diagnosing psychiatric conditions, establishing credit risk, and identifying the threat of criminal recidivism in parole settings, to name just a few. In good part, we have Paul Meehl’s early work to thank for these achievements.

At the same time, any epistemology that takes science seriously, must note the similarity between the considerations made by contemporary epistemologists and those made using clinical approaches in psychology. They engage in many of the same cognitive activities that make clinical judgment demonstrably less reliable than actuarial judgment. Traditional epistemologists tend to require that certain conditions of justification be met before a belief meets the standard of knowledge. Satisfaction of this requirement is usually documented non-quantitatively, and without regard to either the cost associated with establishing justification or the

importance of the issue addressed. This may be OK for theoretical physics, but theoretical physics does not claim to be a normative endeavor—to offer humans useful guidance about important matters. Epistemology does. If it is to deliver on its normative promise, traditional epistemology must embrace the findings of (what might be called) Ameliorative Psychology, and start charting the course of epistemic excellence: The efficient allocation of cognitive resources to robustly reliable reasoning strategies, all applied to significant problems. This approach combines at least economics, psychology, and philosophy, creating the kind of rich intellectual gumbo that naturally occurring theoretical questions produce. It is fitting that the philosopher-psychologist Meehl should have done psychological research that would have so potent an impact in a central area of philosophy.

The contributions of Meehl’s approach are at least as substantial in practical matters that have intellectual impact. Reporting statistical power in journal articles is now much more common, and there is at least pause in the blithe reporting of *P*-values. In philosophy of science, test severity is now more commonly evaluated in terms of whether the background theory is hard or soft. To a first approximation, to say that a theory is hard is at least to say that the largest part of a theory’s most important theoretical claims are accurate. Consisting of so many accurate claims, a hard area of psychology makes it easier to identify the suspect claim when the hypothesis fails a test. As a result, hard psychology can far more easily construct severe tests than soft psychology. There may be other features distinguishing hard and soft theories, but this one is the most important consequence of Meehl’s conception of psychology.

One final contribution of Meehl’s work to philosophy is worth mentioning, because it has influenced a kind of naturalism in the philosophy of science of which he not only approved but did much to fashion. Philosophers of science who use the history of science as evidence often did so by studying selected cases. But the representativeness of these cases is uncontrolled, and so subject to biases about the relative frequency with which a piece of evidence would occur independent of the philosophical hypothesis proposed. This observation by Meehl amounted to an embarrassment to philosophers of science, particularly those trained in quantitative methods, and with the chagrin came at least some reforms.

Many of these themes were present in Meehl’s early work, but they were consolidated in “Theoretical Risks and Tabular Asterisks”. What makes Meehl’s achievements so unusual is that he secured them in the face of so stiff a headwind. His contentions seldom enjoyed the easy trajectory of fashion, and their consequences were usually disruptive of received practice and sometimes even humiliating to well-meaning researchers. So, it is a testament to his persistence and clarity of vision—perhaps even the accuracy of his view—that his work is now so widely influential. In the course of a career, one can perhaps hope to make a novel discovery, to work out a position in satisfying detail, or even to influence a

generation in one's field. Meehl realized this hope several times over.

References

- Bishop, M., & Trout, J. D. (in press). *Epistemology and the psychology of human judgment*. New York: Oxford University Press.
- Godfrey-Smith, P. (2003). *Theory and reality: An introduction to the philosophy of science*. Chicago: University of Chicago Press.
- Meehl, P. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46*(4), 806–834.
- Trout, J. D. (1998). *Measuring the intentional world: Realism, naturalism, and quantitative methods in the behavioral sciences*. New York: Oxford University Press.
- Trout, J. D. (1999). Measured realism and statistical inference: An explanation for the fast progress of 'hard' psychology. *Philosophy of Science, 66* (Proceedings), S260–S272.