First published in *Journal of Clinical Psychology*, 2005, *61*, 1317–1354. Extract, pp. 1339–1345.

Meehl on Metatheory

Disagreements

Some mild demurrers aside, this review of Meehlian[1] metatheory has so far been commendational. But there are large gaps in Meehl's evolving outlook, comparable to drafting the body sculpture and interior accoutrements for an advanced automotive design while neglecting to allocate space for motor and fuel. The aesthetic features needn't be incompatible with the overlooked power components, but until those are also worked into production schematics the company's body shop had better hold back on cutting and casting.

Meehl's major metheoretic omissions, the residue of Popperian thinking, are twofold:

a) his corroboration (crypto-confirmation) is indiscriminately holistic, and

b) he seemingly ignores scientific discovery.

Let's start with "discovery", mainly as commonsense understands this but also as a theme in the philosophy of rational belief (cf. Reichenbach's famous contexts of discovery vs. justification). Both Popper and Meehl of course appreciated that in order to test hypotheses one must first obtain hypotheses to test. But neither, so far as I can find, published anything probative about the outset epistemic status of those. What I find on "discovery" by word search in Meehl's published articles is mainly reference (notably 1990a, p. 33; 1990b, p. 137; 1992b, pp. 134, 160, 163, 167) to Reichenbach's distinction between the "context of discovery" and "context of justification," about which he says nothing beyond advising retention of some updated version thereof, and mention of discovery in his own research. Also, Meehl (1992a) speaks repeatedly of discovery as a normal scientific activity. But in his unpublished (1990b) he approved of discovery most unequivocally in comment on Watson & Crick's famous DNA finding: "The example also shows how Popper, Reichenbach, and the Vienna positivists were wrong in saying there could be no logic of discovery (despite Popper's title)" (1990b, p. 25). In contrast, Popper's position on discovery was hard-core negative: Despite the tin-ear translation of his 'Logik der Forschung' booktitle as 'The Logic of Scientific Discovery' (its last word should have been 'inquiry' or 'research'), he seems to have rejected altogether the possibility that a theory might have some epistemic merit prior to testing. (Cf.

[1] (Ed.) Paul E. Meehl (1920–2003), clinical psychologist and philosopher of science, whose work WR greatly admired.

"The initial stage, the act of conceiving or inventing a theory, seems to me neither to call for logical analysis nor to be susceptible of it"—Popper, 1959, p. 31. Since this act perforce incorporates some degree of uncertain belief, Popper presumably excluded this outset belief from the reach of normative appraisal as well.

Yet if no conjectures can warrant appreciable credence prior to testing, then neither should our fallible beliefs in test outcomes be warranted until those in turn are tested, and so on into vicious regress.[2] So a comprehensive account of warranted belief must include some views on the epistemic management of opinions acquired or modified in ways other than hypothesis tests. And indeed, one variety of acquiring hypotheses by discovery has long been both a mainstay paradigm of learning from experience and a classic philosophic conundrum of justification to which Meehl repeatedly alludes as "Hume's problem," namely statistical induction. This reasons that when almost all the N things of kind K observed so far have had property P, it's quite likely, if N is large, that almost all kind-K things, or at least all that we encounter subsequently, will also have P. And although the weight of evidence accumulating when Ks are encountered singly in sequence could be viewed as concatenated corroboration from implicit repetitious testing of "Most Ks have P" even when this generalization doesn't consciously occur to us until N is quite large, we would surely attain much the same confidence in this same inductive conclusion from initially finding a large flock of Ks wherein the strong prevalence of P-ness elicits our first opinion on P's incidence among Ks. The metatheoretic point to be taken here is that garden-variety statistical induction is not a specialized form of hypothesis testing. Rather, the generalities it yields are driven from the outset by data that shape their propositional contents even while conferring plausibility upon them. It is, in short, a primitive version of *discovering* generalities that seem lawful. But from the long history of philosophers' failing to justify this pattern of inference (never mind that this failure could alternatively be taken to discredit their standards of "justification"), Popper concluded as preface to his hypothesis-testing model of progressive science that a rational "inductive logic" does not exist. One might counter that taking statistical induction's rationality to be impugned just by the inability of philosophers to justify it would be not just commonsensically absurd, but lethal if taken seriously. But that is too simplistic a rejoinder: The philosophic issue has been not whether we should continue this inferential practice but whether meta-reasons can be developed for doing so. Yet it is meta-irrational to insist that no proffered explanans[3] should satisfy us unless it too has been explained, albeit neither should we foreclose the

[2] This is not to insist that data beliefs need justifying in precisely the same way that hypotheses of the sort Popper wants justified need this. Rather, it submits that we are not entitled to posit a sharp divide between these absent a plausible epistemic argument for that.

[3] "Explanans": (a) An assertion proffered to explain something; or (b) the state of affairs so asserted.

possibility of deeper explanation.[4] In his published work, Meehl stayed well clear of induction's justificational quicksand, but in (1990b), unpublished, he repeatedly stepped to its edge and refused to be sucked in.

Whatever its justification, classic statistical induction is just the simplest form of our inferring generalizable explanations of observed events from our discovery of structure in data collections. Peirce's label 'abduction' for inferences of this sort has finally begun to achieve some popularity, though promiscuous usage is impairing its value for metatheoretic discourse. Closer to home, I have for quite some time repeatedly argued that inferences from newly observed data patterns to lawful explanations created for them, which I originally called 'ontological induction' but have since relabeled 'explanatory induction', are prevalent both in technical science and everyday life. I will not here develop this thesis yet again; should you care, you can find a decently nontechnical exposition with additional references in Rozeboom (1997, pp. 366–389). These afford only the opening chapters on explanatory inductions (EI for short), whose different forms taken in different situations are surely more variegated than the ones I have explicitly identified. And although it would please me if EI could entirely replace hypothesis testing in our metatheoretic recipes for scientific theory development, I have little doubt that in epistemic practice there will always be theoretical terrain that EI cannot invade until astute leaps of imagination bring back scouting reports authenticated by hypothesis tests *à la* Meehl. Even so, if only my voice could carry like Meehl's, I would insist that EI be given equal billing with hypothesis testing in our graduate methodology education. I regret that Meehl didn't pick up on this issue when we could have had some instructive debate on it.

Actually, Meehl was involved in discovery-oriented theory development throughout his career, starting with his research on the MMPI[5] and acknowledged in his latter-day side remark, 'I believe strongly in "exploratory" and "refined folk-observational" knowledge' (1990b, p. 173). In Meehl (1978), he gave some specifics of parameter estimation in his own research practice, which he also cited more generically in his explicit metatheoretic framework for theory development diagramed in his (1990b, p. 116) Figure 2. The basic point to take on this is that EI comprises *means* of theory development, not alternatives to it. Parameter estimation is not an immediately evident instance, but neither is it plain how that fits into the hypothetico-deductive model of theory adjudication. Indeed:

> How can hypotheses be tested by predictions that derive from or explicitly incorporate the values of parameters that are open (that is, unspecified) in the hypothesis tested?

[4] How deeply should we attempt to explain the cogency of induction? I can't say, but here's a comparable ontological issue: Why does the universe exist? If we argue that God (or something akin thereto) created it, then how do we explain the existence of God?

[5] (Ed.) The Minnesota Multiphasic Personality Inventory, a well-known clinical psychological test that Meehl helped to develop.

If a hypothesis H needs specified parameters to entail a crucial testable prediction when H itself does not fix those, how do we get the parameter values to test?

Although Meehl did not address these questions explicitly, his implicit answer to the first was including function forms in his list of things a theory might predict (1990a, p. 130). Inferring just the form of a function relating specified variables, or more generally a pattern over an ensemble of dataset properties (notably in modern multivariate analysis, an array of covariances or other joint-distribution moments) is a prediction that existentially quantifies over the ranges of open parameters in that pattern's description; and in principal (though never exactly in noisy practice) the test data will identify those parameters if that prediction is true, or disprove it otherwise. And in answer to the second question, the fitted parameter values can then be used to strengthen the prior theory by specifying smaller windows of uncertainty for those. Note, however, that each tightening—in Meehl's corroboration formula, decreasing the width and perhaps placement of the interval within which a numeric prediction receives full C-credit—is a discovery-induced change in the theory tested, not continued testing of the one corroborated previously. This procedure is not an unrealized ideal, but is true of real-life pattern fits which, when overdetermined as required to earn respect, are never exact but only trends within a scatter of approximation errors, uniquely defined only relative to a more-or-less arbitrary fit measure. And the salient point to take from this is that these parameter estimates are not Popperian free-style speculations but directed discoveries.

When a nascent theory has open parameters, it is difficult if not impossible to find a graceful way by which these can become specified under the hypothesis-testing rubric for theory development. Consider the following challenge to hypothesis testing's alleged superiority over inductive discovery in this case: Suppose that you have access to a large database (many thousands of human subjects) with observations on many items of personal information (medical assessments, sociological and genetic characteristics, scores on the items in aptitude and achievement scales, etc., details of which don't matter here). And you have also conceived a novel theory T_0 of human development which implies that certain parameters ϕ of these items' joint distribution in the unbounded population of which your observed subjects are a finite sample should be appreciably non-null. (ϕ comprises, say, certain special contrasts in this score distribution, or open parameters in a structural model thereof for which this data configuration enables a determinate solution."Null" is a baseline expectation, *inter alia* zero for relational and contrast measures, and Normal for higher distribution moments.) Moreover, you are not content merely to support T_0 by establishing a few tiny departures from Null among the ϕ parameters, but seek strong corroboration of T_0's strengthening to a T_i that specifies each parameter by an interval whose thickness (width) is negligi-

ble. How can you corroborate thick predictions of these parameters—*persistently* corroborate, not just mix hits and misses—even as you revise T_0 to shrink their widths?

One way to do this (best? only?) is through a series $S_1, S_2, \ldots, S_i, \ldots, S_n$ of increasingly large samples drawn randomly without replacement from your database. The scores observed in each S_i yield a sample estimate $\hat{\phi}$ of the population ϕ-values together with an appraisal of their uncertainty, which advise you to replace T_i by a T_{i+1} positing updated ϕ-values from which you then predict ϕ in S_{i+1} less thickly than in S_i. (One good way to update the prediction intervals is by making them high-p confidence zones estimated for ϕ from all the previous samples combined. Or if for some reason that seems illegitimate, you could use just the information in S_i.) If you pace this series astutely with a very large size of final S_n, you can expect strong corroboration of your final T_n's thin prediction of S_n's ϕ-values. And depending on how you think corroboration of one theory rubs off on others similar to it, T_n should inherit some accumulated corroboration from the prediction performances of prior T_i in the series as well. If you do admit corroboration transfer among similar theories, you may want your number of steps leading to T_n to be rather large, since that gives you many corroborations to combine. Otherwise, n needn't be larger than 2.

All this is very well: We can indeed modify each T_i's parameter conjecture as current data sampling advises and corroborate the improvement in a new sample. But how might choosing n to be 2 or more yield an epistemically firmer conclusion than just $n = 1$, that is, simply estimating ϕ's component values by tight intervals (sampling-theoretic confidence zones or, if you prefer, some alternative expression of residual uncertainty) obtained from the full database without making any prior predictions of their values? If we arrive at the same data-driven theory at end, why should it matter if any preceding corroborated predictions have been taken from it? At least in this case (not in all, but that's a larger story), outset induction from the full database, leaving no fragment of that behind on which to attempt corroboration, surely yields fully as much support for T_n as does some tortuous sequence of partial-data corroborations. I'm not suggesting that corroboration is unimportant. Obviously a theory's track record matters when we contemplate gambling on its implications that absent the theory are still uncertain. And seeking to test a novel prediction can lead us to abductively provocative observations that we would never have stumbled upon absent that guidance. But I do submit that if a data finding D urged by theory T does not impart credence to T regardless of whether we were antecedently aware that T predicted it, D's epistemic support for T is illusory. Debate, anyone?

That prior interest in certain theoretic possibilities may have motivated search for those no more disqualifies their claim to discovery status than a mineralogical prospector's find of a valuable ore deposit doesn't really count as "discovery"

because he was looking for something like that. In both cases the searcher could have stumbled on this find without a search plan (arguably this is how most commonsense dispositional attributes become conjectured). And also in both cases, although the prospector may well be in an uncommon situation (controlled data-harvest or geographic traverse, respectively) deliberately contrived to promote possible manifestation of something that only a specialist in this matter could recognize or even conceive, what he finds can differ so much from what he was seeking that he abandons the quest that brought him there to explore the unexpected discovery instead. Thus when harvested data show not the pattern anticipated but conspicuous manifestation of something quite different, EI may well proffer a skeletal explanation for that which, in its prospect for confirmation and elaboration by ensuing research, does more to advance our comprehension of the phenomenon at issue than would an estimate of parameters in the outset model.

Also deflecting my scowl at Meehl's metatheoretic neglect of scientific discovery is his repeated commendation of "convergent lines of evidence" (e.g. 1990b, p. 118) and Salmonish "damned strange coincidence". Discovering that certain pattern features of data collectable in a controlled observational setting (the sorts of abstracta that "parameters" characterize) systematically recur (approximately) over multiple sectors or aspects of the data structure is typical of the adventitious input to which EI is responsive; and further discovery of interdependencies among which features recur under what conditions puts EI into powerdrive. (Were I to flesh out this grandly schematic claim with some examples, I would start by pointing out features common to all pairs of points in a bivariate numeric distribution wherein linear regression has zero residual scatter, and move on to patterning that can be found in the observations afforded to students in an introductory chemistry lab.) It was only Meehl's metatheory, not his scientific practice, that neglected discovery.

Even so, it was a serious deficiency for Meehl to have omitted any articulate endorsement of discovery from his didactic on science's epistemic endeavors. Possibly he felt that this was so thoroughly embedded in scientific practice that it didn't need any metatheoretic defense. Unhappily, that is not so: In at least some sectors of behavioral science today (just how pervasively I am not qualified to say), the simplistic Popperian model of theory development sets the standards for publication acceptability and students' research-methods education. This is especially true of structural equations modeling (SEM), which is the approach to analysis of multivariate covariance data that has largely superceded its exploratory factor analysis (EFA) precursor. From what I discern from monitoring the SEMNET listserver traffic, the following admonitions to SEM neophytes are only mild parody of attitudes that currently prevail among its dedicated partisans:

1. Since SEM's state of the art affords no advice or traditions for creating hypotheses whose confirmation would enhance our understanding of the events

modeled, feel free to let the SEM algebra and solution procedures most familiar to you guide and constrain your creation of causal-path hypotheses to test.

2. Feel no obligation to generate or search out data for analysis that overdetermine your model's solution beyond the bare minimum required for a unique solution. In particular, to avoid needless risk of model misfit, include at most three indicators for any latent variable you hypothesize.

3. Your modeling results are not worth publication unless your goodness of model fit passes a statistical hypothesis test at an orthodox alpha level. And if your solution fails this test, you must not submit a fit to these data made acceptable by revising your model constraints. No post-hoc model fit, no matter how tight, gives any probative support to a hypothesis educed from the errors in the data's reproduction by a less successful model.

4. Your statistical test of model fit is indifferent to what population is sampled by your data so long as the sampling has been suitably random. So to avoid setting unwanted precedents, be reticent when professing to identify this population. And if your fit is successful, don't waste time and risk confusion by voicing concern for whether the substantive nature of the latent variables implicated by your data might differ from the interpretation you have antecedently posited for them. Your model has passed its significance test and that's all SEM standards require for confirmation of your tested hypothesis.

Although assent to these norms for SEM practice is reassuringly far from universal among its practitioners, I sense that students of multivariate methodology are being indoctrinated so to much the same degree as they have, at least until recently, been put in thrall to NHST. Hence if respect for Meehl's metatheoretic stature can be transformed into educational import, SEM instructors should be urged to read Meehl on both statistical testing and verisimilitude. Although SEM's significance testing is the "strong use" that Meehl approved (cf. 1990b, p. 116f.; 1997, p .407f.), it is far *too* strong; for Meehl insisted that we must also allow interval predictions, and the notion that a path model should be rejected just because one or even many of the pathweights and residual covariances it posits to be zero aren't *exactly* so must have seemed as absurd to Meehl as it does to me. Even more saliently, Meehl's latter-day push to give regimented confirmation credit to near-miss test results puts him in direct conflict with intolerance for model solutions that fall a little short of an arbitrary standard of near-perfection. And surely Meehl would have been appalled at admonitions against modifying one's analysis of a given dataset in light of an instructive modeling failure, albeit Meehl's own failure to publish his reasoned views on scientific discovery deprives us of appeal to his authority on this point.

The other huge omission in Meehl's metatheory, this one unredeemed and foundational, lies in its treating the confirmation resulting from a test of theory T as change in the credibility of T's entirety with scarcely any manifest concern, except when trying to divert blame for test failure from a favored theory's core, for how that distributes differentially over the ensemble of propositions collected in T. When T implies both D and E, where D is a test prediction and E is some other entailment of T (E could be a core postulate in T, or some conjunct in T's auxiliary hypotheses, or an additional observation-language consequence of T, or a large chunk of T selected for special interest), verifying or refuting D does not in general corroborate E to the same degree or even the same direction as it does for T as a whole. As Meehl himself repeatedly emphasized, this is obvious when D proves false, since even though that falsifies T and every other theory/hypothesis/conjecture that also implies D, T is generally a conjunctive composite of many propositions (technically, T can be viewed as equivalent to the conjunction of all propositions entailed by it or, restricting this to what we can actually verbalize, to any finite truncation thereof that entails the rest), and only one of those components needs be false to discredit T as a whole. So to grasp the full epistemic import of T's D-test misfire, we should try to discern the credibilistic impact of *Not-D* on each conjunct in T. (This recognition is automatic in the Bayesian model of rational belief change, except that its principle cannot be practiced due to insufficient identification of the relevant prior and conditional credibilities.) Of course we can't do it all, at least not explicitly; but we can and should attempt to search out and appraise those components of T that seem most salient for what we want to do about this discrediting of T. Above all, if we have been partisans of T we can hope to salvage what we find attractive in its core by altering dubious presumptions in its auxiliaries. Meehl took pains to recognize this strategy of theory repair (cf. 1990b, p. 121f., so although by rights he should have said more about distributing blame on the downside of test outcomes (cores can't be protected come what may), that merits only a critical frown. But failure to allocate differential credit for a test's *success* is quite another matter.

When T entails D, failure of D's disproof to discredit every propositional part P of T has a mirror image in failure of verifying D to confirm every P in T. Indeed, when D and E are both deductive consequences of T, verification of D *may* also confirm E—which I submit is prevailingly presumptive, else why should we be so willing to trust new predictions from a theory whose previous predictions have all proved successful?—but plainly does not always do so. One construction showing this is E = *'Either not-D or T'* which, when T entails D, is another deductive consequence of T such that verifying D confirms T but decreases the plausibility of E unless *not-D* was certain at outset. And if that construction is too artificial to trouble you, here is an importantly realistic eruption of this epistemic problem:

> Advanced theories in the same real-world area of application often agree in some predictions while disagreeing on others. Suppose that T_1 and T_2 both entail D for a test not yet undertaken while T_1 also entails an additional prospect E, logically independent of D, that T_2 strongly disputes by entailing $\sim E$. (E vs. $\sim E$ may emerge from elaboration of inconsistently different positions on a controversial uncertainty which is not directly testable because E or its denial is part of its respective theory's nonobservational core. Or E could be the possible outcome of a crucial experiment that, technically or financially, is not yet feasible.) If testing verifies D, this confirms both T_1 and T_2. But since E and $\sim E$ are mutually exclusive and jointly exhaustive, any increase in the credibility of one must in a rational belief system be accompanied by decrease in credibility of the other. The only rational alternative to one of T_1 or T_2 having its stand on E disconfirmed by its success at predicting D is for the credibility of E and hence $\sim E$ to remain unchanged by D's verification.

The point to be taken here is that Meehl's righteous condemnation of statistical null-hypothesis testing's most egregious blunder, thinking that confirmation of a statistical hypothesis similarly confirms the substantive theory from which that was derived, likewise applies to unthinking generalization of a theory's confirmation by a successful test thereof to increased confidence in other implications of the theory. Some of those are indeed confirmed thereby, but others are not and may even be disconfirmed at least a little. So verifying a theoretical prediction is—or should be—only the first phase of extracting the epistemic import of this test result for the theory at issue.

References

Meehl, P. E. (1990a). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant using it. *Psychological Inquiry*, *1*, 108–141, 173–180.

Meehl, P. E. (1990b). *Corroboration and verisimilitude: Against Lakatos' "sheer leap of faith".* Working Paper MCPS-90-01. Minneapolis: Center for Philosophy and Science, University of Minnesota.

Meehl, P. E. (1992a). Factors and taxa, traits and types, differences of degree and differences in kind. *Journal of Personality*, *60*, 117–174.

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.

Meehl, P. E. (1997). The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* Mahwah, NJ: Erlbaum.

Popper, K. R. (1959). *The logic of scientific discovery.* London: Hutchinson & Co.

Rozeboom, W. W. (1997). Good science is abductive, not hypothetico-deductive. In L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* New Jersey: Erlbaum.