

November, 2005

## Meehl on Metatheory\*

Wm. W. Rozeboom  
University of Alberta

*Abstract:* A critical survey of almost everything Meehl ever wrote on the philosophy of scientific research.

\*I want to acknowledge my debt and gratitude to W. M. Grove, Meehl's close friend and colleague, for his cogent referee commentary that helped this biographic safari to stay on track of the big game.

Invitation to contribute a memorial overview of Paul Meehl's contributions to psychonomic metatheory has been an honor I could not decline. But attempting to do even scant justice to this theme has proved to be a humbling experience. I have, of course, long appreciated Meehl's stature as our discipline's foremost concept analyst: My first conceptual publication, written in graduate school, was a defense of MacCorquodale & Meehl's (1948) prevailingly misunderstood distinction between intervening variables and hypothetical concepts; and although my thesis research on the behavior-theoretic "What is learned?" issue was designed before I learned of Meehl's work on latent learning, his 1954 monograph with MacCorquodale on Tolmanian expectancy theory -- by an order of magnitude the best work of psychonomic concept explication to have then appeared -- reached me in time to enhance my already-strong desire for a postdoc at the Univ. of Minnesota's Center for the Philosophy of Science where Meehl was then an active participant. My fellowship application was successful, yielding two years in the ferment of Feigl's group that benefitted me enormously. But Meehl had discontinued his active Center participation shortly before my arrival, and to my great loss and regret I never managed to bring about any personal interaction with him then or later. It was clear to me early on that Meehl was special. But just how awesomely special I did not fully appreciate until my recent immersion in his full publication output that includes a richly informative autobiography (Meehl, 1989). The latter reveals him to have been extraordinarily gifted in both his intellectual endowment and, despite certain personal tragedies on which he does not dwell, an early environment that nurtured its maturation. It is rare for an adolescent to have ready access to college-level literature on psychology, statistics, law, logic and analytic philosophy, much less the motivation and ability to devour this with comprehension. Just as extraordinary was having a gang of friends (his "Young Logician's Group" -- 1989, p.342) whose version of work-up baseball was turns at pitching arguments for randomly assigned positions on religious/political controversies with failures in a pitcher's argument to withstand reasoned criticism scored as runs-against. (The recreational-sports metaphor is mine, not Meehl's.) And at his Minneapolis doorstep was a major-league university availing psychologists and analytic philosophers at the leading edges of their respective specialties with whom he was quick to establish rapport when only a college undergraduate. The wealth of productions that ensued from this unique confluence of interests, abilities, and opportunities is what we celebrate here. Meehl's autobiography makes abundantly plain that intelligent thinking -- its pursuit, appreciation,

and practice -- was the dominant force in his life. And intelligence for him wasn't just acquisition and retention of diverse archived information (though he wasn't averse to flaunting his superiority at that in his writings) but, more important, proficiency in articulating the deeper complexities of concepts and issues while giving all diverse outlooks on these a fair but critical hearing. So it is no surprise to find that Meehl was much concerned throughout his career with the theory and praxis of effective reasoning. Much of this was abstract metatheory on "hypothesis testing" which he developed as an admittedly revisionist Popperian and on which this commentary will eventually focus. But unlike most of us in academia, Meehl also shouldered responsibility for theoretical inferences having real-world consequences. Mainly these were personality appraisals, creating/adjudicating/applying concepts of enduring human attributes that importantly influence what people feel and do. And although his initial MMPI work thereon in collaboration with Hathaway was comparable to an industrial engineer's crafting prototypes from design schematics, he soon became a practicing psychotherapist as well, applying theorized personality constructs to assess his clients and guide their therapy. I am not qualified to reflect on Meehl's contributions to clinical psychology; I merely note that his theorizing had consequences in the real lives of real people for whom he undertook professional responsibilities. Neither can I judge the extent to which the feedback this gave him on the practical payoff of these constructs influenced the development of his metatheoretic views; but there must surely have been an effect which future exegesis on Meehl's legacy may want to explore.

Be that as it may, it is well worth appreciating that Meehl was deeply involved over the full gamut of psychonomic theorizing from hands-on applications to human affairs, through creating and adjudicating his own personality-theoretic hypotheses derived from that experience, to effectively incorporating in his metatheoretical writings (though more as consumer than producer) the most advanced views on the cognitive character of scientific theories developed by modern analytically sophisticated philosophers of science. I find it helpful to overview Meehl's abundant contributions within this arena as development of persistent themes that may be differentiated as

- A. The nature of theoretical constructs.
- B. Statistical diagnosis of clinical categories.
- C. Detection of hypothesized taxa.
- D. Dislodging psychology's Null-Hypothesis Significance Test (NHST) incubus.
- E. Complexified hypothesis testing (confronting auxiliary hypotheses and verisimilitude).
- F. Cliometric theory appraisal (a late-career reincarnation of theme B).

This listing fails to record that Meehl's superb ability to recognize, articulate, and clarify confoundings, hidden complexities, and other underappreciated obscurities in our important working concepts (I'm pretty good at this myself, so believe me when I tip my hat to the master) also graced contributions to religion and jurisprudence

(Meehl et al. 1958; Meehl, 1959; Meehl, 1965; Livermore & Meehl, 1967; Meehl, 1971). But these are not plainly thematic, so I shall say no more about them here. Neither shall I say much about themes B, C, and F inasmuch as others in this symposium are far more qualified than I to do so. But A and D need overview here despite their formerly hot issues for psychonomic science having become largely quiescent in recent years. And I shall also argue that Meehl's development of E suffers from major omissions.

#### A. Meehl on theoretical constructs.

As old-timers in this venue will recall, psychology's vocalized metatheory during the first half of the 20th Century was strongly polarized not merely over which current theories might be plausible but more fundamentally on whether scientific psychology could ever use mentalistic concepts save as targets of derision. And the semantics of theoretical terms also loomed large then in theoretical physics and advanced philosophy of science, which argued less-than-pellucid theses under the labels "logical positivism" and "operational definitions" that soon became psychonomic buzzwords as well. To overview this issue -- it was important, and remains so even though it should no longer seem threatening -- and Meehl's contribution thereto, I cannot improve on Rozeboom, 1984, excerpted here with some compressions marked by square brackets:

The skeptical temperament that pervades science and analytical philosophy takes these epistemic-engineering problems of explanation very seriously indeed. ... Logical positivism courageously sought to face down this doubt [whether explanatory theories can ever qualify as scientific knowledge] in the only way that classical epistemology could envision, namely, by proposing that theoretical sentences which seem to make claims about unobserved entities do not really do so [but rather, must be translatable into sentences using only our data language if they are to be cognitively meaningful. Operationism supported this positivist thesis by holding that such translations could be contrived from data-language statements of form 'If ... then \_\_\_'. However,] though virtually all theoretical constructs in scientific practice have analytic *if/then* implications, seldom do theoretical meanings seem to be *exhaustible* by a conjunction of test/outcome conditionals. ... But explicit definition is not the only way in which new concepts can be derived from old ones; and as the possibility of less straitened forms of definition became appreciated, logical positivism gave way in the 1950s to the empirical realism long championed by Feigl (cf. 1950, 1956) under the title "logical empiricism." This is the thesis that although theoretical terms get their meanings from the data-language contexts in which they are used, what they semantically *designate* are causal features of natural reality generally concealed from perception but knowable from their data consequences. ...

Within psychology, moreover, contrary to the mythology widely propagated by ignorant or malevolent bystanders, the methodology of scientific constructs practiced by [most] leading behaviorists and other theorists of operationist persuasion ... was an indigenous empirical realism long before philosophers learned how to distinguish this from positivism. When Hull and Tolman sought to develop behavior theory in terms of "intervening variables" they understood this label *not* in the positivistic sense later proposed for it by MacCorquodale & Meehl (1948) but as demarcation for hypothesized causal mediators ("hypothetical constructs" in MacCorquodale & Meehl's sense). [See especially Hull 1943 pp.21-23 and Tolman 1959 pp.97f.]

In explicit opposition to positivist doctrine, the view that psychology's theoretical constructs designate real underlying causes through their conceptual roles in a "nomological network" was forcefully articulated by Cronbach and Meehl (1955). And continued efforts by psychologists to pin down what they meant by *operationism* (which few comprehended in the philosophers' narrowly technical sense) eventually clarified this, too, as our version of empirical realism. [See especially Garner, Hake, and Eriksen, 1956, p.158; also Campbell and Fisk, 1959 p. 101, and Campbell, 1959 p.175ff.]

The brevity of reference to Meehl's work in this abridged overview does inadequate justice to its psychonomic impact at the time. The paper with MacCorquodale evoked considerable response which, though of mostly dubious quality, served at least to make serious metatheoretic debate welcome in our literature.<sup>1</sup> And his later paper with Cronbach empowered psychometric assessment with the advanced understanding of theoretical concepts' nature and validation developed by Feigl and his collaborators. This introduced "construct validity" into APA's official Test Standards manual as commended concern for the accuracy with which test instruments diagnose non-manifest attributes thought to underlie overt behavior. Although Meehl never published any distinctive views of his own on the semantics of theoretical constructs, he continued throughout his career to endorse the philosophically sophisticated view of "open" concepts and implicit definition. (See Meehl, 1978, p.815; 1990a, p.240f.; 2002, 361f.; 2004 online, p.8].)

#### B. Meehl on clinical diagnosis by statistical inference.

His early interest in clinical psychology maturing into hands-on psychiatric practice, conjoined with his extensive engagement in psychometric development of quantitative instruments to assess clinically significant personality traits, uniquely positioned Meehl to write THE book (Meehl, 1954) on tough-minded inference applied to tender-minded practicalities. Despite the galvanizing effect this seems to have had in its milieu, debate on the actuarial approach to clinical diagnosis was already well-mounted and Meehl did not in this work advance its applied methodology. What he did accomplish, for any reader who could follow carefully articulated explanations, was expunging large conceptual confusions that were apparently degrading the use of quantitative methods in clinical practice. And in so doing, he powerfully demonstrated that James' (1897) famous tough-minded/tender-minded division of intellectual styles is not strictly dichotomous despite often appearing so: Meehl was as tough-minded as we come. Yet he had so receptive an appreciation for the riches in (some) tender-minded notions that he repeatedly (not just in this instance) felt it worth his while to clear away the obscurities that so often degrade those, like cleansing classic artworks of accumulated grunge to bring out their underlying

---

<sup>1</sup> Confusions that prevailed in this literature were unfortunately exacerbated by M&M's choice of 'hypothetical constructs' to denote what Tolman and Hull had called 'intervening variables'. In his autobiography (1989, p.371), Meehl rues unintended connotations of the noun in 'hypothetical construct' but doesn't acknowledge that M&M's redirection of 'intervening variable' caused much greater mischief.

elegance. Theme B in Meehl's career is not a foreground concern of this essay, but one way or another it is background to those that are.

### C. Meehl on taxon detection.

During the late '50s (cf. Meehl, 1989, p.365), Meehl developed an interest in the roots of schizophrenia that infused his work for the remainder of his career. First came his theory of manifest schizophrenia's etiology (Meehl, 1962) positing that this originates in a single defective dominant gene -- a "schizogene" (Meehl, 1972, fig.1) -- which causes development of a neurological disorder ("schizotaxia") whose afflicttees "become, on all actually existing social learning regimes, schizotypic in personality organization" which in turn disposes development of overt schizophrenia in a minority of cases having "other ... constitutional weaknesses and put on a bad regime by schizophrenogenic mothers" (1962, p.[143 in Psychodiagnosis: Selected Papers]). Or as he put it in 1989 when still unhappy with how firmly he had established this, "we believe [that we can soon] definitively corroborate or refute my conjecture that schizophrenia is a low-probability ( $p < .10$ ) decompensation of a soft neurological integrative disorder (schizotaxia) which is inherited as an autosomal dominant of 100 percent penetrance." (1989, p.376).

Meehl's dissatisfaction voiced in this autumnal quotation was no way incurred by want of testing. For quite some time he had sought to support his schizogene theory with hard quantitative evidence; and not finding available techniques of data analysis entirely suited to his need, undertook to develop his own. As fate would have it, this was an ideal challenge for Meehl to demonstrate that he could practice the metascience he preached -- defining theoretical constructs as nodes in a nomological net (Feigl) on one hand, and theory appraisal by strong empirical tests (Popper) on the other. And although the logic of fitting nomological nets to multivariate quantitative data had already become familiar in the burgeoning literature on mathematical models, it seems that his distrust of how suited those were to the distributional structure posited by his conjecture -- categorical source attributes ("taxa") working with manifold additional influences to produce observable outputs most usefully idealized as gradations on metricized continua -- motivated him to develop an approach of his own. The foreground task was to infer, from subjects' ratings on three or more numerical scaled indicators taken in clinical practice to diagnose schizophrenia, whether the joint distribution of scores on these indicators support the hypothesis that these express, albeit with questionable reliability, an underlying common-source variable having just two levels, schizotypic vs. non-schizotypic; and if so, to infer the proportion of subjects in each taxon as well as a subject's taxon probabilities given his indicator scores. From there, Meehl's schizogene theory becomes testable from schizotype concordance among the close relatives of schizotypic probands.

The source-structure presuppositions on which Meehl grounded his original MAXCOV procedure for such taxon diagnosis are essentially those invoked by Lazarfeld's prior Latent Structure Analysis (LSA), which

likewise solves for source categories albeit from data that are also categorical in contrast to Meehl's metric data dimensions. Meehl could have used LSA to test for schizotype taxa simply by treating score intervals on his indicators as data categories; but possibly that loss of information would have incurred solutions inferior to what MAXCOV obtained. Be that as it may, Meehl pursued MAXCOV over many publications thereafter, most recently in Waller & Meehl (1998), which expands its range of applications while also recognizing its proximity to other name-branded procedures for latent-source recovery. Appraising MAXCOV's method prowess or past applications payoff is not on this essay's agenda. But I will return to schizotype detection when considering Meehl's more recent views on theory testing (goodby Popper, hello Lakatos).

Not all tests of schizogene theory proffered by Meehl employ MAXCOV. A simpler one that will later be useful as foil for voicing reservations about his most recent account of theory testing was published in Golden & Meehl (1978), with an overview in Meehl (1978), p.820. The most immediate problem for researching schizophrenia under Meehl's theory thereof is how to identify people who are schizotypic (the near-certain developmental consequence of schizotaxia) even in the "compensated" majority thereof whose schizophrenic dispositions remain unrealized. Several indicators of compensated schizotaxia are known, but each has imperfect accuracy and it is clinically important to estimate the true-positive and false-negative conditional probabilities for these diagnostic instruments.<sup>2</sup> Meehl points out how to do so for a given binary indicator  $X$  of schizotypia when we have identified a large group of decompensated schizotypic individuals ("probands") whose immediate family and close relatives are also available and willing to be assessed for schizotypia. Given some help from auxiliary distribution idealizations, schizogene theory authorizes derivation, from the marginal and joint frequencies of this symptom in these probands' parent pairs, the probabilities  $Pr(s | X+)$  and  $Pr(s | X-)$  that a person who respectively does or does not manifest  $X+$  is schizotypic. And from there the idealized theory further deduces the incidence (relative frequency) of symptom  $X+$  in various inheritance classes of these probands' near relatives -- monozygotic vs. dizygotic twins, other siblings, grandparents, etc. Meehl viewed checking the accuracy of such outreach predictions following parameter estimation as "consistency testing", a strand of his metatheory on theory appraisal that will resurface below. Rather more can be inferred from schizogene theory, creating additional tests thereof, when covariances among two or more schizotype indicators augment the analysis; and more yet when MAXCOV is brought to bear on joint distributions of graded schizotypia indicators. But that's deeper into Meehl's taxon detection than we need to go here.

---

<sup>2</sup> For the Meehl-Golden tests of schizogene theory, each indicator is taken to have binary (yes/no) output. But most of these are cuts on more finely graded rating scales, so an even more basic applications issue is where to put the cuts. That is the problem which Meehl designed MAXCOV to solve.

#### D. Meehl on Null-Hypothesis Significance Testing.

Meehl was not the first psychologist to decry our long-institutionalized tradition of appraising research results by null-hypothesis tests, but by a large margin he was our most persistent scold. Probably best known is his first assault on this (Meehl, 1967); but his (1978, 1990*a*, 1990*c*, 1997) and (2004 online) all contain extensive critiques of NHST's idiocies. How much corrective influence, if any, these had is hard to discern; Meehl himself sensed little and took personal affront from that. (See Meehl 1990*c*, p.108; 2004 online, p.5.) Even so, he enriched the debate over NHST with three salient points that would merit thoughtful discussion even were our research community's peer-pressured<sup>3</sup> addiction to null-hypothesis testing in full remission.

The first, which he credits to Bolles (1962) but surely appreciated on his own, is need to distinguish the statistical hypothesis  $H$  directly tested by NHST from the substantive theory  $T$  from which  $H$  is derived only with the aid of implausibly idealized presumptions about how a study's observational setup distributes probabilities over its possible outcomes. Unlike Fisher's NHST-originating agricultural research (cf. Meehl, 1990*a*, p.226*f*.) or more generally in what Meehl (1997, p.394*ff*.) describes as "technological" (vs. "theoretical") contexts, there has been only a loose cognitive coupling in the behavioral sciences between substantive theories and their statistical shadows in dataspace; and we still have neither reasoned metatheory nor reliable intuitions about how sampling inferences about population statistics should instruct our opinions about why these have the values they do.

Second, we have Meehl to thank(?) for emphasizing the "crud factor", a suitably honorific label he credits to Lykken for the pervasive interconnectedness of things whereby it is virtually certain that any two passively observed variables will correlate at least modestly. Or as Meehl puts it as a quoted truism, 'In the social sciences and arguably in the biological sciences, "everything correlates to some extent with everything else".' (1990*a*, p.204) He persistently argued that this correlational background noise results in our journals becoming glutted with experimental tests of provocative theories sustained by their failing to disconfirm at low alpha levels even while also failing to show effects strong enough to credential any as winners.

Third, and most salient from Meehl's "somewhat Popperian" (1967, p.112) perspective on theory appraisal, was his distinction between "strong use of significance tests" (his wording) vs. "weak use" thereof. Strong testing requires a strong theory to test, and Meehl's paradigm for those are advanced physical theories which imply, among other theory-specific features listed in Meehl (1990*c*, p.114. fig.1), precise or narrowly

---

<sup>3</sup> I would like to think that most traditional null-hypothesis testing has resulted from real or expected social coercions (demanding advisors, opinionated editors), like driving on the right to avoid the grief we would otherwise get from others. But in (1990*a*), p.237, Meehl had a more contemptuous suggestion: "Since the null-hypothesis refutation racket is "steady work", and has the merits of an automated research-grinding device, scholars who are pardonably devoted to making more money and keeping their jobs ..... are unlikely to contemplate with equanimity a criticism that says that their whole procedure is feckless and that they should quit doing it and do something else."

bounded values for certain summary measures computed from data collected under suitably prepared experimental conditions. Bearing in mind that a NHST's "null" value can be any value of the test statistic on which we choose to center our data report, we are to view use of NHST for testing a substantive hypothesis  $T$  as "strong" if we infer from  $T$  the statistical hypothesis  $H$  that our test statistic's population value  $\theta$  should be null, or is "weak" if our  $H$  inferred from  $T$  is only that  $\theta$  is not null. These are extremities of strong vs. weak use; more generally, as illustrated by one-tailed  $H$ -tests, inferring  $\theta$  to lie in a restricted interval is intermediately weak or strong according to this interval's width in comparison to  $\theta$ 's full range of possibilities. And theory  $T$ 's strength regarding  $\theta$  is graded according to how narrow an interval  $T$  authorizes for an  $H$ -test of  $\theta$ .

Meehl did not describe NHST testing of point predictions as "strong use" until his 1990c paper, by which time he had become unenthusiastic about all versions of traditional significance testing, even confidence intervals, embarking instead upon an even stronger approach to theory testing to be discussed next. But the weak-test absurdity of professing to support a favored theory by inferring from it a non-null effect of unspecified strength whose apparent NHST success could just as well be manifestation of the background correlation noise was his outset case (Meehl, 1967) against NHST albeit crud by that name did not surface until Meehl (1990a). Unfortunately, his extensively referenced (1967) exposition is much weaker than his powerful but seldom cited putdowns of NHST in (1990a,c), each of which rehashed the crud menace but made its modest relevance more clear than did (1967).) It is surely true that almost every correlation or effect size in any psychonomic database would prove to differ "significantly" from zero in a sufficiently large subject sample, whence the official goal of two-tailed NHST is to answer with some statistical uncertainty a question which was never plausibly in doubt. (A  $p$ -value exceeding our chosen alpha signals in practice not the absence of a conjectured effect, but only that our sample size was too small to detect it.) But why should that matter? Suppose that a major advance in quantum physics implies that in your research area, effects detectable by NHST at  $\alpha$ -level .01 only in subject samples larger than 10,000 cannot exist, so that a sufficiently large sample can be decisive. Would that weaken NHST's absurdity for psychonomic research? Not at all for two reasons, one of which is Meehl's own knockout punch and another which he ignored.

The point that Meehl never mentioned is that an NHST's chosen significance level has little relevance to what we really do in response to the data this appraises. Officially, Types 1 vs. 2 errors are responding incorrectly when the hypothesis tested is respectively true or false. But our actual response is *never* unconditional execution of a binary accept- $H$  /reject- $H$  action choice as NHST pretends. We don't even know what it means to "accept" or "reject" a tested  $H$ . (If you think it might be deciding for sure whether  $H$  is true, think again. You can choose whether to profess what you believe but not whether to believe what you profess.) What we really get from an

NHST report is a sense of this putative effect's strength, an update of motivation to continue this line of research, and maybe some leverage for soliciting publication and funding. Crud's prevalence doesn't matter for any of that.

And neither does it much matter for Meehl's lethal point, which addresses use of NHST to appraise substantive hypotheses. He acknowledges that in applied research, whose primary concern is whether controllable input to our subjects appreciably affects their resultant attributes, the statistical question that NHST seeks to answer is pretty much our target except that we want to know not just *whether* but *how strong*. (Problems with crud still arise here, but can be largely suppressed by good experimental design.) But in "soft psychology" (Meehl's apt locution), the statistics tested for NHST significance generally measure interdependencies among output variables and are of interest not for their own sake but as extrusions from a theory *T* of the underlying states/processes which give rise to these dependent observables. When *T* implies (tightly or loosely as may be) a testable statistical hypothesis *H*, refutation of *H* shows that either *T* itself or inferring *H* from it can't be altogether correct. But conversely, verifying *H* does little to increase *T*'s plausibility unless we can largely rule out possible reasons for this effect alternative to what *T* posits. Chances are that it's not hard to think of other possibilities. But we don't need crud to discourage the leap from establishing *H* to confidence in *T* -- it suffices simply to appreciate that *T* is not the only conceivable source of the *H*-effect. Indeed, one might wonder if Meehl wouldn't have made his case against NHST more sharply concise if he hadn't wanted an excuse to report the plethora of crud connections he and Lykken had found in the responses of 57,000 Minnesota highschool seniors to several dozen sociological questions (1990a, p.205f.). Or maybe he assumed, perhaps correctly, that few of his readers would accept that a ritual so entrenched as NHST could be shown feckless by just a few well-chosen words. And it may also have been his style impediment at work. For Meehl did have a style defect. He always had so much to say, and so much savored saying it, that more often than not he embellished the meat of his message with needless authority references, fulsome examples, and fringe associations by which most readers must surely have been awed but also, I venture, overloaded. In the present instance, he could have destroyed respect for NHST as a research tool almost by just two short passages here condensed from three at source -- "almost", because some exegesis helps to make their point clear:

1) ... there is a negligible difference between the substantive theory and the counter null hypothesis in agronomy, whereas in theoretical psychology they are distinctly different and frequently separated by what one could call a large "logical distance"; (1990a, p.224) "... the substantive theory has a host of alternatives." (1990a, p.229)

2) "One reason why psychologists in the soft areas naively think that they have strongly proved a weak theory by a few significant [test statistics] is that in their education they learned to conflate *statistical significance* with the broader concept of *evidentiary support*. So they are tempted to believe that if there is

nothing wrong with the experimental design, or in the choice of statistic used to test significance, they are "safe" in concluding for the verisimilitude of a theory. Pedagogically, I have found the quickest way to dispel that comforting illusion is to put the question, 'Assume you *had* the parameter; what would you know, and how confidently?'" (1990*c*, p.117). [Only the last sentence in this quote really matters; I include the rest to clarify that Meehl's "comforting illusion" is confusion of support for  $H$  with similarly strong support for  $T$ .]

To backstop Meehl's case, consider interrogating as follows the prime mover of an impending research project to appraise the still-unknown value of statistical parameter  $\theta$  under conditions  $C$ :

a) Do you want to know  $\theta$  for its own sake, or only because learning this is inferentially relevant to something else less directly observable? If for its own sake, say because knowing  $\theta$  will advise you how strongly and/or reliably a controllable manipulation brings about a certain desired result, or tells how accurately some data variables in your study can be predicted from others, does it suffice just to learn that  $\theta$  does NOT have some particular value, say zero, under conditions  $C$ ? (This might be enough if you are merely doing a pilot study to convince others that this line of research merits support, in which case you may well not want to learn too much about  $\theta$  too quickly.) Or wouldn't you rather learn *what* value  $\theta$  has under  $C$ , or at least roughly so, even if you have no immediate need for tight precision. If the latter, you want an estimate of  $\theta$ 's value together, preferably, with some appraisal of its accuracy; and in that case, whereas NHST is then worthless for you, a confidence interval derived from your data by the very same statistical logic as NHST and almost the same computations tells you essentially all that you can learn about  $\theta$ -in- $C$  from these data unless closer study of their distribution casts doubt on their compliance with your statistical model's premises.

b) If what you learn about  $\theta$  is intended to inform us about something else as well, what is that something else, and by what line of reasoning do you go there? In particular, if you hope to adjudicate some favored theory  $T$  that professes to explain, at least in part, why  $\theta$  has the value it does in  $C$ , what are alternatives to  $T$  that would also explain this -- if you don't know of any, is that because there are none or only because you haven't tried to think of others? -- and what can any experiment tell you about  $\theta$  that supports  $T$  more than its competitors? If you have been so preoccupied with testing  $\theta$  that you haven't considered this deeper question, take time *now* to think on what you should do next if you do manage to ascertain the point value of  $\theta$  in  $C$ , or nearly so. If you haven't a clue, don't waste your time and others' resources on  $\theta$ -focused research.

And where does this leave crud prospects? Not as a threat but an inducement: If we seek we will surely find. It assures us that if we become intrigued by a new flavor, facet, oddity, or other neglected aspect of researchable events, a sufficiently large collection of multivariate observations thereon will exhibit relationships that NHST vouchsafes real enough to encourage sponsored continuation of your inquiry. The real problem is one that, when I was young and pets ran loose, was shared by the neighborhood dogs who loved to chase moving cars: If you catch one, do you know what to do with it?

To summarize we have a minor point leading to a major one. Important only as a historical footnote is that Meehl's crud-based arguments against NHST were largely digressive. It should have been obvious, though apparently was not, that trying to learn a statistic's population value by appraising just whether its sample value is nonzero, or only that it differs in a chosen direction from a chosen target value, is astonishingly benighted when confidence intervals also afforded by the standard statistical model grounding NHST tell that as a fragment of so much more. And since psychology's methodology establishment (notably journal editors and graduate text authors) seemed unable to comprehend the naked logic of this or admit to practicing *statisticus interruptus*, Meehl may have hoped that wrapping this in threats of crud contamination would promote better statistical hygiene. But at end Meehl (2004, p.5) conceded failure. Even so, like honeybees to flowers, our research community would surely have swarmed out to harvest point estimates of its statistical parameters had only metatheory been in hand for processing the nectar of high-quality statistics into the honey of substantive explanations thereof. And that is the major conclusion to take here; that what our research methodology still sorely lacks and from which NHST has been a blighting distraction, are recognized techniques for learning not just *what* values our data's statistical parameters have, but *why*.

#### E. Meehl on post-Popperian hypothesis testing.

The philosophers with whom Meehl interacted so extensively at the Minnesota Center for Philosophy of Science during his early years had no practical advice to offer on interpreting statistical parameters either. Although modern analytic philosophy has made sophisticated advances on some fronts, nondemonstrative inference is not one of them.<sup>4</sup> Meehl first voiced his outlook on hypothesis appraisal during his initial attack on NHST:

... inadequate appreciation of the extreme weakness of the test to which a substantive theory  $T$  is subjected by merely predicting a directional statistical difference  $\bar{d} > 0$  is then

---

<sup>4</sup> Mid-century analytic philosophy did indeed grapple mightily with the character of confirmation, but these struggles were mainly attempts to get leverage on its conceptual nature, notably by Carnap (1962) and Hempel (1945) ( who both had ties to Feigl's Center), and produced nothing of value – or so I submit though Meehl might have disagreed – for assisting research scientists to theorize more effectively.

compounded by a truly remarkable failure to recognize the logical asymmetry between, on the one hand, (formally invalid) "confirmation" of a theory via affirming the consequent in an argument of form:  $[T \supset H_1, H_1, \text{infer } T]$  and on the other hand the deductively tight *refutation* of the theory *modus tollens* by a falsified prediction, the logical form being:  $[T \supset H_1, \sim H_1, \text{infer } \sim T]$ .

While my own philosophical predilections are somewhat Popperian, I daresay any reader will agree that no full-fledged Popperian philosophy of science is presupposed in what I have just said. The destruction of a theory *modus tollens* is, after all, a matter of deductive logic; whereas that the "confirmation" of a theory by its making successful predictions involves a much weaker kind of inference. This much would be conceded by even the most anti-Popperian "inductivist". The writing of behavior scientists often reads as though they assumed -- what is hard to believe anyone would explicitly assert if challenged -- that successful and unsuccessful predictions are practically on all fours in arguing for and against a substantive theory. (1967, p.112)

Meehl's argument here is unassailable: If theory  $T$  entails statistical claim  $H$ , disproving  $H$  also disproves  $T$  whereas verifying  $H$  does *not*, conversely, verify  $T$ . But unless  $H$  is certain at outset, its verification must enhance  $T$ 's credibility to some extent in a coherent belief system -- just how much or, antecedent to that, by what conceptual resources can we say how much, remaining large epistemic questions not here addressed. That is no objection to Meehl's argument, just observing this to be only a small fragment of an account of rationality barely launched. It is also to note the first appearance of Meehl's viewing confirmation of predictions as weak whereas refutation thereof is strong, albeit his double quotes around 'confirmation' acknowledges Popper's disapproval of this concept.

A decade later, we get a considerably more elaborate model of theory appraisal, still basically Popperian but with more degrees of freedom. After speaking in strongly pejorative terms about psychology's infatuation with refuting null-hypotheses, he launches his alternative:

It is easiest to see this [null-hypothesis foolishness] from the methodological viewpoint of Sir Karl Popper, but fortunately we have here a rare instance in which Sir Karl's position yields the same result as the Bayesians, and both give the same result as "scientific common sense" practiced by those chemists and biologists who know nothing about philosophy of science or Bayesian statistics and could not care less about either. Briefly and simplistically, the position of Popper and the neo-Popperians is that we do not "induce" scientific theories by some kind of straightforward upward seepage from the clearly observed facts, nor do we "confirm" theories as the Vienna positivists supposed. All we can do is to subject theories--including the wildest and

"unsupported" armchair conjectures (for a Popperian, completely kosher) -- to grave danger of refutation in accordance with the formally valid fourth figure of the implicative syllogism  $p \supset q, \sim q, \therefore \sim p$ , Popper's famous *modus tollens*. A theory is corroborated to the extent that we have subjected it to such risky tests, the more dangerous tests it has survived the better corroborated it is. (Meehl, 1978, p.817)

'Corroborate' is the verb Popper choose to mean "increase the credibility of" because he mistakenly took 'confirm', our most familiar word for that, to connote the certainty of 'verify' (Popper, 1959, Ch.X[p.251], fn.1), and hence disallowed that conjectures of the sort considered "theory" by scientists could ever be confirmed. But starting in (1967, p.106) and for the most part thereafter, Meehl seems to have accepted 'confirm' and 'corroborate' as largely equivalent for denoting what positive evidence does for a theory's plausibility. Even so, during a hiatus circa 1990 he took pains to proffer a Popperian definition of corroboration -- "[the theory] has been subjected to a test and has not failed it" (1990c, p.110) -- that makes no reference to credibility. (In contrast, Popper's 1959 Appendix \*ix developed metricized formulas relating a corroboration's strength to the tested theory's logical probabilities before/after the test). Initially I took Meehl's failure to connect corroboration with rational belief change here to be an oddly inept omission; but after some prompted reconsideration I now discern method in this madness. More on this later.

A second Popperian concept that Meehl first deployed in Meehl (1978, p.818ff.) is "verisimilitude", which in first approximation, while acknowledging that the concept still lacked satisfactory explication, he provisionally clarified as "truth-likeness" or "nearness to the truth". And professedly in reaction to critiques of Popper by Lakatos and Feyerabend, though this extension could surely not have been new to his thinking, he expanded the  $[T \supset O, \sim O, \therefore \sim T]$  formula for *modus tollens* refutation of a theory  $T$  by falsifying its observational consequence  $O$  to acknowledge that the theoretic antecedent in a Popperian test is far more logically complex than adequately schematized by a single letter:

... when we spell out in detail the logical structure of what purports to be an observational test of a theoretical conjecture  $T$ , we normally find that we cannot get to an observational statement from  $T$  alone. We require further a set of often complex and problematic auxiliaries  $A$ , plus the empirical realization of certain conditions describing the experimental particulars, commonly labeled collectively as  $C$ . So that the derivation of an observation from a substantive theory  $T$  amounts always to the longer formula  $(T \cdot A \cdot C) \supset O$ , rather than the simplified schema  $(T \supset O)$  ... The *modus tollens* now reads: Since  $(T \cdot A \cdot C) \supset O$ , and we have falsified  $O$  observationally, we have the consequence  $\sim(T \cdot A \cdot C)$ . Unfortunately, this result does not entail

the falsity of  $T$ , the substantive theory of interest, but only the falsity of the conjunction ( $T \cdot A \cdot C$ ); that is, we have proved a disjunction of the falsities of the conjuncts. So the failure to get the expected observation  $O$  proves that  $\sim T \vee \sim A \vee \sim C$ , which is not quite what we would like to show. (Meehl, 1978, p.818)

How we do show what we would like to show isn't addressed in Meehl (1978); but that wasn't Meehl's target in this paper and before critiquing his matured views on theory appraisal it's best to finish inventory of his specialist (neoPopperian) conceptual tools.

Considerable time elapsed between Meehl 1978 and his next foray into wide-spectrum metatheory; but when he returned in Meehl (1990*a,b,c*) he had much to say. Meehl (1990*a*) was a carpet bombing of pervasive incompetences of design and data interpretation in extant correlational research that could well profit behavioral-science graduate programs to adopt as their main text for a compulsory methods seminar. But it didn't modify Meehl's previous account of theory appraisal beyond a modest inflation of his ' $T \cdot A \cdot C \rightarrow O$ ' schema that Meehl (1990*c*) expands further.

Meehl's last substantial paper on metatheory, (1990*c*), was also in a sense his first inasmuch as he had never previously addressed the praxis of theory development much more comprehensively than urging us to waive null hypotheses in favor of theories yielding precise predictions. (That didn't instruct us *how* to create such theories plausible enough to warrant the expense of testing but, counterfactually, it should at least have goaded NHST partisans to upgrade their research aspirations.) Because this finale opens so many important issues with positions thereon that need dispute, I will quote Meehl (1990*c*) extensively (the indented paragraphs below), with elisions marked "..." and condensations or occasional clarifying insertions by me in square brackets. After some introductory remarks he continues on pp. 109-112:

For ease of reference, I set out the general position with brief numbered paragraphs and minimum elaboration or defense.

1. A scientific theory is a set of statements in general form which are interconnected in the sense that they contain overlapping terms that designate the constructs of the theory. In the nomological network metaphor (Cronbach & Meehl, 1955), the nodes of the net are the theoretical constructs (entities) and the strands of the net are the functional or compositional laws relating them to one another. Contrary to simplistic operationism, it is not required that all the theoretical constructs be operationally defined. Only a proper subset are linked in a direct way to observational predicates or statements. In idealization, the theory consists of a formal calculus and an embedding text that provides the interpretation of expressions in the formalism (cf. Suppe,

1977). The empirical meaning of the theoretical terms is given partly by "upward seepage" from the subset that are operationally tied to the data base. Logicians explicate this upward seepage by means of a technical device called the Ramsey sentence, which eliminates the theoretical terms without "eliminating the theory" or repudiating its existence claims. For psychologists its importance lies more in showing (contrary to simplistic positivism and freshman rhetoric class) how a system of expressions can both *define* and *assert* concurrently. ... [outside references here followed by an interesting list of important nonobservational concepts, notably ones with a causality flavor, that often occur in theories but seem to get their meaning in obscure ways other than] this "implicit definition by Ramsified upward seepage".

2. In conducting an empirical test of a substantive theory  $T$  (which is imperative to distinguish from a test of the statistical hypothesis  $H$ ) the logical form is the following:

$$[f] \quad (T \cdot A_t \cdot C_p \cdot A_i \cdot C_n) \rightarrow (O_1 \supset O_2)$$

where  $T$  is the theory of interest,  $A_t$  the conjunction of auxiliary theories needed to make the derivation to observations go through,  $C_p$  is a *ceteris paribus* clause ("other things being equal"),  $A_i$  is an auxiliary theory regarding instruments, and  $C_n$  is a statement about experimentally realized conditions (particulars). The arrow denotes deduction (entailment), and on the right is a material conditional (horseshoe) which says that if you observe  $O_1$  you will observe  $O_2$ . ( $O_1$  and  $O_2$  are not, of course, related by strict entailment.) On careful examination one always finds in fields like psychology that the auxiliary  $A_t$  is itself a conjunction of several auxiliary theories  $A_1, \dots, A_m$ . If in the laboratory, or in our clinic files, or in our field study, we observe the conjunction ( $O_1 \cdot \sim O_2$ ) which falsifies the right-hand conditional, the left-hand conjunction is falsified *modus tollens*.<sup>5</sup> .....

3. [Here, Meehl points out again (cf. 1990a) that falsifying the observational consequence at right of ' $\supset$ ' in theory-test schema [f] entails falsification only of the conjunction of propositions to its

---

<sup>5</sup> The role of  $O_1$  in [f] is to describe the experimental setup or observational preconditions for determining whether  $O_2$ . But Meehl seems to have overlooked here that on occasions when "in the laboratory, or ... field study" we observe  $\sim O_1$ , which we can easily contrive by botching the setup, this entails ( $O_1 \supset O_2$  regardless of whether  $O_2$  is true and is hence a form-[f] test result that corroborates  $T$ . This intolerable implication spectacularly illustrates why the material implication ' $p \supset q$ ' (by logicians' definition equivalent to 'Either not- $p$  or  $q$ ') is an importantly inadequate substitute for our ordinary-language subjunctive 'If  $p$  then would  $q$ '. Meehl was surely aware of analytic philosophy's extensive literature on this non-equivalence, and he could have avoided the problem it creates here simply by shifting  $O_1$  in [f] into the string of conjuncts on its left; yet his emphasis in Meehl (1997 p.399), on the materiality of this conditional strongly suggests that he wanted it this way. His stated reason is far from adequate (though we will later note a benefit); and this point much merits serious discussion because the material conditional in [f] omits an essential facet of experimental research, namely, that some variant of  $O_1$  is almost always a major part-cause of  $O_2$  if that occurs on the test occasion at issue; and our deepest challenge here is to make that connection more conceptually explicit than accomplished merely by reverting the horseshoe in [f] to a subjunctive conditional.

left, telling us that at least one of those conjuncts must be false but not identifying *which* of them are wrong.]

4. If this falsification does not occur [that is, if  $O_1$  does yield  $O_2$ ], we say that the theory has been *corroborated*, which for Popper means that it has been subjected to a test and has not failed it. Whatever affirmative meaning (reliance? "animal faith"? rational scientific belief?) we give to corroboration derives from a further statement, namely, that absent the theory  $T$ , the antecedent probability of  $O_2$  conditional upon  $O_1$  is "small". If that is not so, our corroboration (pre-Popperians called it *confirmation*, a term that Popper avoids as being justificationist) is weak, some say negligible. ... When we speak of the theory as "taking a risk," as "surmounting a high hurdle," as not being flunked "despite a dangerous test," these locutions refer to the notion that on some basis (prior experience, other theory, or common knowledge and intuition), *absent the theory  $T$  we have our eye on*, we see no reason for thinking that  $O_2$  has a high probability conditional upon  $O_1$ .

5. The obvious way in which we warrant a belief that  $O_2$  has a low prior probability conditional upon  $O_1$  absent the theory is when  $O_2$  refers to a point value, or narrowly defined numerical interval, selected from a wide range of otherwise conceivable values. ...

6. By the instrumental auxiliaries  $A_i$  I mean the accepted theory of devices of *control* ... or of *observation*. In some sciences (e.g. nuclear physics), it would be quite difficult to parse these theoretical claims from the theory being tested, but such is not the case in the behavioral sciences. [Galvanometers and Skinner boxes are cited as examples.] I am using the term ['instrument'] narrowly ... to stipulate that the theory of an instrument must not contain, explicitly or implicitly, any psychological constructs or theories. ...

7. In his discussion of the positive and negative heuristic, Lakatos (1970) lumped all the conjuncts on the left except  $T$  as part of the "protective belt," and maybe even portions of  $T$ . ... Lakatos also subsumed both *disturbing particulars* (one way to violate  $C_p$ ) and *incomplete statement of auxiliary general laws* ( $A_i$ ) into his ceteris paribus clause. I think it is important to distinguish these [for reasons having to do with theory revision].

8. In the presence of what appears to be a falsifying protocol, the Lakatosian methodology prescribes a strategic retreat (a *Lakatosian defense*, I call it). *When* adoption of this strategy is warranted, instead of confessing immediately that  $T$  has been falsified and should be abandoned, [this] remains to be discussed: In what follows immediately I consider the *literal truth of  $T$* , because we can't discuss everything at once. In reality, a sensible psychologist would take it for granted that

$T$  itself is almost certainly imperfect, either in (a) the weak sense that it is *incomplete* or (b) the strong sense that it is, when taken literally, *false*. This involves the problem of verisimilitude, and the important Lakatosian distinction between saying that a theory is falsified and saying that one ought rationally to abandon it. In science, theories when falsified are not abandoned prior to a Kuhnian revolution (Kuhn, 1970), but are appraised as to their degree of verisimilitude, and attempts are made to patch them up. But in discussing the Lakatosian strategy of retreat, I initially set aside the problem of verisimilitude of  $T$  and reason as if we wish to defend it literally as it stands. ...

9. In conducting the strategic retreat in the presence of accepted falsifiers is it useful to think in terms of a theory as attempting to deal with several fact domains. ... [That is, big theories have implications in several nonoverlapping "phenomenological" domains,] and a theory's ability to handle facts in qualitatively diverse domains is more impressive than its only handling a large number of particulars belonging to the same domain. ... Suppose  $T$  is doing very well in several domains, and it has also succeeded with a few high-risk predictions in a subdomain in which also, however, the conjunction  $(T \cdot A_i \cdot C_n)$  has been clearly falsified. Then an obvious strategy is to amend the domain  $C_n$ . ...

10. A related situation exists with regard to the theoretical auxiliaries  $A_i$  where one asks how widely  $A_i$  is found in the various derivation chains in different domains before modifying it to deal with a subdomain falsification. A further criterion is the extent to which a certain auxiliary has been independently corroborated in other experiments not involving the  $T$  of current interest. ...

11. This strategic retreat--beginning with the reluctant admission of the falsifying protocol, then cooking up new auxiliaries by denial of the ceteris paribus clause in troublesome domains, and then challenging some of the former auxiliaries themselves--may finally result in recognizing that the program begins to look somewhat "degenerate," as Lakatos called it. ... [(Remainder omitted here speaks inconclusively about ad-hocery and when to challenge the theory's hard core.)]

12. Like the concept of verisimilitude, the metaconcept of core or central portions of a theory has not been given rigorous definition, and I am not able to offer one. It is obvious that some such distinction must, however loosely, be made. Intuitively one sees that in a particular theory some components are ubiquitous in dealing with the range of facts whereas others are not thus centrally located, although they are truly "part of the theory". .... [(Elided is Meehl's partial explicating what parts of a theory qualify as its "core". He considered that to have some importance

in appraising the theory's "verisimilitude", a concept he introduced next; but details on core don't matter for what ensues here.)]

Still more Meehlisms need inspection before I submit a major disagreement with his metatheory's epistemic orientation; but some qualms also warranted by Meehl's numbered positions above can best be expressed now, before additional complexities intrude. To begin, after initiating his Thesis 1 with a useful and almost correct review of modern philosophy's post-Positivist understanding of theoretical constructs in science (see theme A, above), Meehl contaminates this with an atypically maladroit introduction to Ramsey sentences. I shaded "correct" with "almost" here because, although Meehl's metaphor of theoretical and observational concepts tied together as nodes in a net of conjectured laws is apt, describing this process of implicit definition as "upward seepage" is rather less so. For one, the connotations of "seepage" (*slow* and *shapeless*) traduce the brisk edification this explication of concept development affords. But it is Meehl's "implicit definition by Ramsified upward seepage" phrase, echoing the defect in his description of what a theory's Ramsey sentence is, that demands correction. For what is at issue here is central to what has been 20th Century philosophy's stormiest controversy, namely the epistemic character of theoretical concepts. Should you wish to follow Meehl in certain passages of his most advanced metatheorizing (notably 1990*b* and 2002*b*) you will need to get clear on this. Nor should you want to shun it. If the connotatively sterile label "Ramsey sentence" puts you off, be apprised that *prime consequence* (of the theory from which it derives) could be a replacement name for this that advertises its epistemic import.

Strictly speaking, no theory-explicating Ramsey sentences exist in real life, since these are linguistic idealities. They presuppose theories expressed as a finite set of fully explicit declarative sentences in a formalized language (*a*) having precise inference rules that capture what educated commonsense accepts as valid deduction, and (*b*) whose descriptive terms (expressions that under a semantics for this language profess to have referents) have been partitioned into two disjoint subsets that we can label "theoretical" and "observational", respectively. (The Ramsey construction is indifferent to how that dichotomy is imposed, and different applications are free to choose whatever cut is most illuminating for the analysis at hand. Most appropriate to treat as "observational" when considering alternative theoretic explanations for a given research finding would be the leanest vocabulary adequate to describe the empirical results plus an insuppressible minimum of causal-cohesion notions such as listed in Meehl (1990*c*, 2<sup>nd</sup> col., lines 19ff.) But don't let the artificiality of this setup turn you away: Real-language counterparts of Ramsification have major import for real-world theory adjudications, and the idealized formalization lets you see clearly what might well be obscure in a theory's ordinary-language expression.

Suppose then, suppressing niceties of notation distinguishing between use and mention of locutions, that  $P(a)$  and  $Q$  are arbitrarily complex sentences wherein the predicate  $P(\_)$  ascribed to nominal  $a$  is formalized to contain no occurrences of  $a$ . And let  $(\exists x)P(x)$  be the sentence asserting that there exists something that satisfies

predicate  $P(\_)$ . Letting ' $\rightarrow$ ' abbreviate 'logically entails', it is then straightforward to prove that under standard principles of deductive logic, if  $P(a) \rightarrow Q$  when  $Q$  does not contain  $a$ , then also  $(\exists x)P(x) \rightarrow Q$ . By iteration of this principle, it follows that if  $T(S)$  is a more-or-less complex theory that entails an observation-language prediction  $O$ , while  $S = \langle s_1, \dots, s_n \rangle$  is an  $n$ -tuple comprising the nominal expressions in  $T(S)$  that do not occur in  $O$ , then  $O$  is also entailed by  $(\exists X)T(X)$ , where  $X = \langle x_1, \dots, x_n \rangle$  is an  $n$ -tuple of logical variables of appropriate logical types.  $(\exists X)T(X)$  asserts simply that there is some  $n$ -tuple of entities that satisfy  $n$ -adic predicate  $T(\_)$  without picking out any specific one of these. And when  $S$  comprises all the non-observational terms in  $T(S)$  (its "theoretical constructs"),  $(\exists X)T(X)$  is theory  $T(S)$ 's Ramsey sentence.<sup>6</sup> Every observation-language consequence of  $T(S)$  can be deduced from its Ramsey sentence with total disregard for the theoretical terms in  $T(S)$  -- which points out that these terms most certainly do NOT get meaning, by seepage or otherwise, from  $T(S)$ 's Ramsification, else that same meaning should similarly be conferred on all other tuples of hitherto meaningless terms not in  $(\exists X)T(X)$ . Rather, as Meehl correctly observes, Ramsey sentences are a way to theorize without ever introducing theoretical terms. What does give meaning to the  $S$ -terms is taking theory  $T(S)$  to be an "implicit" definition of the  $s_i$  in  $S$  which tacitly stipulates that these respectively denote the entities in whatever ordered set thereof satisfies predicate  $T(\_)$  -- presuming of course, as the theory's Ramsey sentence asserts, that such an  $n$ -tuple does in fact exist. This linguistic gambit is altogether commonplace in everyday life, two recent examples of which are "Green River killer" and "Washington sniper": Each of these designates the perpetrator of a murder spree through its implicit definition by an evolving predication that initially described just portions of its referential target's antisocial actions but, with increasing elaborations thereof in light of accumulating data, eventually enabled apprehension of an individual satisfying that complex predicate. Note, however, that the Washington sniper proved to be not one person but two. And that is a complication generic to the semantics of implicit definienda or, as Meehl liked to say, "open" concepts: When we say that the  $n$ -tuple  $S$  of theoretical entities named in a true theory  $T(S)$  is whatever ordered set satisfies predicate  $T(\_)$ , we seldom have good reason to posit, more strongly than what the theory's Ramsey sentence asserts, that only one tuple of entities satisfies  $T(\_)$  -- whence it would seem that if the nominals in  $S$  are assigned referents by acceptance of theory  $T(S)$  when  $T(\_)$  has more than one satisfier, then  $S$  must equally refer to all of those. But how can that be allowed, when philosophy of language has always presumed reference (designation, aboutness) to be a many-*one* relation of concepts to objects?<sup>7</sup> Making provision for multiple reference first in semantics and thereafter epistemology is a major undertaking which modern philosophy has yet to shoulder.<sup>8</sup>

---

<sup>6</sup> Named after Frank P. Ramsey (1903-1930), the wunderkind Cambridge philosopher who first proposed this construction.

<sup>7</sup> Ambiguity is not an exception: A word or phrase that is ambiguous evokes different meanings depending on circumstances and sometimes evokes these simultaneously. But those meanings themselves, if referential, are presumed by orthodox philosophy of science to have unique referents.

<sup>8</sup> My own efforts to incite some action on this front many years ago were thoroughly unsuccessful. Rozeboom 1970b may well be too ponderous to encourage casual perusal (though surely no worse than par for this course); but Rozeboom 1960a, part I, is a brief and easy read in ordinary language that makes the semantic contrempts undeniably plain. If you are at all interested in the character of word/world relations, try it; you should be intrigued.

Issues of theoretic reference are important, and will resurface below. But first come some concerns with the Popperian vocabulary that Meehl still favored for his hypothesis-appraisal metatheory despite repeatedly professing apostasy from Popperian dogma. In overview, his thesis is that the primary way to advance scientific knowledge -- for Popper, the only way -- is to speculate profusely (not everything at once, but to keep it coming) on what might possibly account for the observables in a given field of scientific inquiry, and then winnow down the hypothesized alternatives to a precious few by evaluating their success or failure at falsifiable predictions. Meehl initially (1978, 1990*a,c*) called these performance challenges "risky tests", but later (Meehl, 1997) reconceived them as "theory appraisals" to avoid needless Popperian taint -- wisely, since both adjective and noun in the earlier label proffered nits to pick. But he never wavered from Popper's approved label for what theories gain from verified predictions, namely, "corroboration". This is one entry into a convoluted cluster of English locutions pertaining to belief and veridicality -- e.g., 'assurance', 'certainty', 'confidence', 'conjecture', 'contention', 'conviction', 'suspicion', 'credence', 'faith', 'truthful', 'plausible', 'doubtful', 'presumptive', and many more -- whose main dimensions comprise differences on one hand in the strengths/intensities/degrees with which people in fact believe diverse propositions, and on the other in the strengths/intensities/degrees with which they *ought* to do so in light of whatever we consider judiciously relevant to that. As a first-approximation -- and we've never gotten much beyond this -- we presume (*a*) that people *do*, in fact, entertain propositions at different levels on a scalable belief-dimension of intentional modality (see Rozeboom, 1972, p.37ff.); (*b*) that it is useful to idealize the contents of a person's graded beliefs as including all propositions that can be expressed in his language, each of which at any given moment is believed -- either occurrently or dispositionally -- at some strength between utter disbelief and complete conviction, and (*c*) that a person's belief system is ideally rational only if, when the atomic propositions therein are treated as a measurable set with its connectors and quantifiers appropriately modeled as set-theoretic relations (inclusion, intersection, union, etc.), his belief-strengths -- I prefer to call these 'credences' though '(personal) probability' is more common in the literature -- satisfy the axioms of probability theory. No real human could ever be even roughly compliant with this idealized model; but it is heuristically useful to pretend that we do, or at least ought to, while additionally conforming for the most part to Bayesian diachronic coherence (cf. Rozeboom, 1997, p.340). Details or development of this ideal aren't needed here, especially since Meehl's own published views on rational belief change, though apparently appreciative of its Bayesian model, barely nibbled at its edges.

The exegetical challenge engendered by Meehl's persistent use of 'corrobor...' as the stem of a transitive verb relating theories to data is to clarify what must obtain for this relation to hold, starting with the ontology of its relata. It does seem reasonably clear that for Meehl, when theory *T* is corroborated by evidence *E*, *T* is a cluster of propositions which for simplicity we can treat as a single conjunction. But is *E* a state of reality (e.g. a cluster of events or some molar aspect thereof), or one or more propositions (or sentences which express them) alleging such a reality? Meehl often speaks of data and "experimental findings" as corroborators or refuters (plainly, whatever

can corroborate one theory might refute another) but that leaves ambiguous whether "data" and "findings" are external-world states of affairs, someone's true beliefs about external reality, or an amalgam of a researcher's true beliefs with the objectivities they are about. I don't know how to resolve this question in the abstract, but here is a thought-problem to bring its issues into sharp focus: Suppose theory *T* entails that if a certain experiment is conducted, a certain gauge will come to register a certain startling value. And imagine two unusual scenarios how this experiment might end:

(a) The gauge does indeed register the startling value, but is destroyed by an explosion (which might also have been predicted had *T*'s implications been thought through more thoroughly) before anyone could observe its reading.

(b) The gauge doesn't show the *T*-entailed value, but blemishes on its dial cause it to be misread as being in the *T*-zone and an untrained lab assistant dismantles the setup before it can be re-examined. (A variant of this scenario could be any sequence of events that causes the gauge to seemingly or actually reach the *T*-zone by some process that overrides the mechanism posited by *T*.)

In case (a), is theory *T* corroborated by this gauge display just prior to destruction -- by *this* event, mind you, not whether a later repetition of the experiment does so -- even though no one knows it? If corroborators are states of affairs independent of whether anyone knows them, then this is surely a corroboration; yet I would not consider it so and unless someone else can make a case to the contrary I shall presume that neither would have Meehl. Here's another twist: Suppose that this experiment yields a totally unanticipated additional effect (e.g. the gauge, whose time-course was recorded to measure response lag, oscillated wildly for several seconds before stopping at the *T*-entailed value), but after much reappraisal of theory *T* it is learned that this too could have been predicted from *T*. Is this post-hoc evaluation at all corroboratory? I recall (though I haven't managed to retrieve the passage) that Meehl did indeed consider such postdictions to be positively probative -- he was far too sensible to think otherwise. But if so, does this additional corroboration occur (i) when the unexpected phenomenon occurs, (ii) when it is recorded, (iii) when that record is first perceived and judged puzzling, or (iv) later when *T*'s entailment of this phenomenon is discerned? I intuit (iv) and will venture this to be consensual.

As for scenario (b), whose analysis I must forego here, this shows (unless your intuitions conflict with mine) that prima facie corroboration of *T* occurs just when someone detects that an otherwise uncertain *E* inferred from *T* is apparently true, albeit violation of background preconditions can dismiss the apparent corroboration as spurious. And the conclusion I take from these thought experiments is that corroboration of *T* by *E* is a *process* wherein some individual *i*'s becoming confident of *E* changes some relation between *T* and *i*. When I am *i*, this *T*'s salient relation to me is its location in my space of propositional attitudes; and I insist that in me, at least, corroborative change is a shift of the corroborated proposition's location on the credence dimension thereof. But Meehl seems never to have made clear whether he would have agreed; and while word search for 'corrobor...' in his papers finds many passages that suggest it, close reading of some, notably the "whatever affirmative meaning ..." waffle in his (1990c) position statement No.4 quoted above, suggests that he wanted to view rational belief change as a consequence of corroboration rather than its constitution. This freed him to propose an innovative

quantitative measure of corroboration while evading the hassle of undertaking an account of how, beyond qualitative increment, a successful corroboration should rationally change one's graded confidence in the tested theory. But it also puts him at odds with what I submit is the sensible reaction to probes (*a,b*) above. Whatever the right of this may be, the issue is provocative and while Meehl can no longer participate, I invite others to pick up for him.

Indeed, Meehl's provocation on this may be considerably more radical than so far acknowledged here. (I have Grove to thank for holding my feet to the fire on this.) He may have sought a notion of "corroboration" that breaks free not merely of Popper's conception but of all traditional intuitions about credence and rational belief change. To pursue this lead we need to bring into play Meehl's other new construct conjugate with reconceived corroboration, namely, 'verisimilitude', which he first used lightly in (1967) without definition, later clarified as "nearness to the truth" (1978,p.818, rephrased as "truth-likeness" in (1990c,p.112), and deployed as an aspirant quantifiable metatheoretic measure in (1990*a,b,c*). I find this word infelicitous both in its syllabicity and its ordinary-English sense of truth-semblance distinct from truth-proximity; but that doesn't disqualify it for silent reading in Meehl's technical sense.

By "verisimilitude", Meehl wanted us to recognize that scientific theories are never impeccably veridical in all respects, whence practical theory adjudication needs to ask not whether true but how so and to what degree. Fully aware of how impossible would be an assessment of theory verisimilitude that specifies all graded ways in which a theory might be approximately true (not even the classic use/mention model of binary truth,<sup>9</sup> which Meehl wisely didn't try to exploit, survives close scrutiny), Meehl nevertheless undertook ground-breaking efforts in that direction.<sup>10</sup> He starts in (1990*c*) by observing that if we have two theories  $T_i$  and  $T_j$  dealing with the same domain and explicitly verbalized in all their cognitively salient respects, notably features such as listed in (1990*c*) Fig.1 (repeated with mild revision in his 1990*b*, Fig.1), we can quantify the similarity of these theories on each of these feature dimensions and thereafter define a composite of those ratings to serve as a global measure -- call it  $\gamma$  -- of similarity between  $T_i$  and  $T_j$ . (Meehl understandably avoided details of the component measures constituting this  $\gamma$ , though he did offer ideas on where to go.) Given this (conjectured) formula for comparing two humanly fallible theories, we then invoke Wilfred Sellars' conceit of Omniscient Jones, an imaginary supreme intellect who knows everything and can verbalize it, let  $T_{OJ}$  be Jones' true theory of  $T_i$ 's domain (the 'oJ' subscript is Meehl's; he failed to anticipate what it would connote in later years), and take  $\gamma(T_i, T_{OJ})$  to be our measure of  $T_i$ 's verisimilitude. Obviously Meehl wasn't expecting any numeric  $\gamma$ -ratings to be made in our real world's near future; but he quite rightly considered this to be a bold conceptual venture at putting substance into the otherwise near-shapeless notion of a theory's proximity to hard-core truth.

---

<sup>9</sup> Namely, that any declarative sentence '*d*' is true iff *d*. The sentence just proffered does not literally assert the principle, but shows schematically what philosophy of language can state explicitly only with the technical contrivance of disquotation marks (notably, Quine's "corners").

<sup>10</sup> One can acknowledge with respect that ground has been broken without agreeing that the ground broken is fertile.

Yet if a theory  $T$ 's position in a graded and apparently multidimensional verisimilitude space is to supersede  $T$ 's binary truth status as our target of  $T$ -appraisal, what happens to the prospect that what we generally get from researching  $T$  is a wobbly level of confirmation, credibility, or probability of  $T$  that eventually approaches an extreme? Classic one-dimensional conceptions of epistemic stature no longer seem adequate, which may well be why Meehl largely evaded speaking so. But are we better off seeking corroboration of verisimilitude? First of all, measure  $\gamma(T_i, T_{OJ})$  of theory  $T_i$ 's verisimilarity to  $T_{OJ}$  is not just incomputable but seriously ill-conceived: If OJ were really omniscient, his language would profoundly differ from ours, not merely in reach but also in precision. (Even if no other issues of shared conceptions intrude, how might OJ match the vagueness that pervades even our most refined concepts?) Whether some subset of his vernacular could be coordinated closely enough with the words wherein  $T_i$  is couched to make  $\gamma(T_i, T_{OJ})$  well-defined is an intriguing question which, however, is best not made pivotal to real-world theory adjudication. Actually, this obscurity arises much closer to home: If theories  $T_i$  and  $T_j$  address the same empirical domain but develop conflicting accounts thereof, can the theoretical concepts implicitly defined by  $T_i$  be matched with those in  $T_j$  closely enough to vouchsafe the feature comparisons on which Meehl proposes to rate their similarity? I haven't studied Meehl's Figure-1 features deeply enough to judge whether intertheory meaning differences degrade comparisons on those, but it's something for future development of Meehl's notions to ponder.

In light of Meehl's superseding Veracity by Verisimilitude, it's not so surprising that he also wanted to shy away from classical confirmation and its cognates. (We can't doubt, or consider plausible, or feel confident of a theory's verisimilitude without some conception of its alternative levels, which Meehl's schematics for  $\gamma(\_, T_{OJ})$  don't provide.) In place thereof, he proffered in (1990b, p.128f.) and (1997, p.415) an index for the corroboration afforded by testing  $T$ 's interval prediction of a one-dimensional numeric parameter  $\theta$  whose value is subsequently estimated empirically. Details of this provisional formula for Meehl's "corroboration index"  $C_i$  (whose subscript indexes order in a series of  $T$  tests) don't really matter here. More interesting is what constituents of corroboration are (a) recognized or (b) neglected by his  $C_i$  measure:

First,  $T$  is permitted to predict merely an interval for  $\theta$ , though the larger the proportion of  $\theta$ 's "Spielraum" (the parameter's range of currently plausible alternatives<sup>11</sup>) excluded by  $T$ 's interval prediction of  $\theta$  (virtually all if a point prediction), the stronger is the test. This exclusion proportion is the theory's "intolerance" ( $In$ ) and was Meehl's proposed value of  $C_i$  for this test when the prediction interval does prove to include the empirical value of  $\theta$ . But unlike Popper, Meehl also endorsed partial credit for near misses. So he took the "closeness",  $cI$ , of  $T$ 's  $\theta$  prediction to be 1 minus the "relative error" (shortest distance of observed  $\theta$  from the predicated interval in proportion to Spielraum), and expanded  $C_i$ 's definition to be intolerance multiplicatively attenuated by suboptimal closeness, that is,  $C_i =_{\text{def}} cI \cdot In$ . Further, corroboration so measured is allowed to aggregate over iterated testing (1990c, p.129f.), enabling the theory to accumulate "money in the bank" if its tests

---

<sup>11</sup> "Spielraum" (play-space) is Meehl's Feiglian label for the target parameter's range of feasibly alternative values. If that isn't logically bounded, its limits are set by current "background knowledge".

are successful. (Provisionally, accumulation sums the single-test  $C_i$  when the tests are "different" experiments but averages them when the tests are replications.)

There is, of course, much in Meehl's nascent corroboration theory to criticize, ranging from (a) technicalities that would be important but straightforward to improve were  $C_i$  to be taken seriously, through (b) forms of testable prediction that aren't plainly focused on numeric data parameters, to (c) conceptual inadequacies that might be lethal. A nice instance of (a) is that, unlike Popperian corroboration which dis corroborates when a theory's test prediction is falsified, Meehl's version doesn't *dis*:  $C_i$  is positive so long as the test statistic is in Spielraum, whereas Spielraum's rationale urges that results falling outside of that are so bizarre that they must surely be *ceteris-paribus* violations and thus don't count against the tested theory. This reluctance of  $C_i$  to disapprove could easily be rectified by rescaling proximity-to-prediction to have a zero point at which scant approval becomes scorn. But lacking some such correction, Meehlian corroboration never makes withdrawals from the theory's ever-increasing "money in the bank". To adopt Popper's brand name "corroboration" for a view that diverges so profoundly from Popper's own, which took refutation to be the driving force of scientific progress, is misleading advertising.

A borderline example of challenge (b) is predicting function forms, albeit those can usually be characterized by special prediction-targetable parameter values (notably zeros) within a broader specified function class. More provocative would be to explore how Meehlian corroboration might track the course of some modest-scope but important real-life theoretic success. A tidy recent instance is whether peptic ulcers might be due mainly to bacteria. When this theory was first proposed by Marshall & Warren in the mid-1980s, it was considered by most medical specialists to be crackpot, yet research tests soon promoted it to a triumphant success (cf. Rogers, A. I., & Hoel, D., 1997). . How might  $C_i$  have been computed, aggregated over these experiments, and thereafter utilized in epistemic appraisal of this theory? The critical data features would have been at least in part (though entirely??) metricized contrasts between experimental and control groups; the conceptual challenge for Meehl's approach, beyond whether measured group differences were all that mattered, is how  $C_i$ -values reflecting the quantified group contrasts would/should be relevant to the impact that these studies would/should have had on medical opinion.

As for problem class (c), it seems strange that Meehl wasn't bothered by the flagrant subjectivities in his  $C_i$  measure, especially if his silence on how a tested theory's  $C_i$ -success should alter its credibility was motivated at least in part by concern for the "sin of psychologism" (Meehl, 1990b, p.32) . Although theories may indeed occasionally entail an interval for some data parameter without favoring any tighter locus therein, surely boundaries for interval predictions of appreciable width would usually be set by someone's personal impressions extrinsic to the tested theory, with the choice of interval unstable not merely between researchers but within as well. (So much for  $C_i$ 's scientific objectivity.) Just as fuzzily idiosyncratic would be Spielraum settings narrower than the tested parameter's logical limits. And does the convenience of algebraic linearity justify Meehl's restricting his corroboration formulas to that form?

Subjectivities aside, how to accumulate corroboration over tests of similar but nonidentical theories is a problem not just for Meehl's  $C^*$  but for all metatheories of progressive theory development. Consider Meehl's schema [f] above, simplified to

$$[f'] \quad T \cdot S \rightarrow E \quad ( S = A_t \cdot C_p \cdot A_i \cdot C_n, \quad E = \text{If } O_1 \text{ then } O_2 )$$

with Meehl's material implication in  $E$  replaced by a more ambiguous ordinary-language conditional to evade the problem cited in footnote #5. What gets corroborated by ascertaining the truth of  $E$  is total theory  $T \cdot S$  comprising the conjunctive entirety to the left of ' $\rightarrow$ ' in [f']. (Meehl makes this clear only for refutation, but gave no reason to think otherwise for successful tests.) Now: If testing  $T \cdot S$  through its prediction [f'] consists of obtaining a data collection that either verifies or refutes complex proposition  $E$ , how is it possible to replicate this test, considering that  $E$  abstracts from specific events at loci (subject times and places) that never recur? Meehl astutely made this possible by putting  $O_1$  to the right of  $\rightarrow$  in [f], rather than on its to-be-tested left as would be the wont of a logician or analytic philosopher, thereby allowing  $T \cdot S$  to include only generalities while sited particulars are supplied as test preconditions in conditional  $E$ . So empirical research can generally contrive for the attribute configuration described by  $O_1$  to be satisfied by different subject samples at diverse times and places, thereby constituting repeated tests of the very same generalized theory  $T \cdot S$ . This also clarifies how it is possible for the same theory to be tested in different ways, namely, by change of  $O_1$ 's predicate in small and sometimes large details (e.g. sample size) from which a corresponding change in  $O_2$ 's predicate yields a test prediction  $E'$  that also follows from  $T \cdot S$ . In such cases it does make sense to envision corroboration of  $T \cdot S$  accumulating over a series of tests even if Meehl's proposed formula for this remains problematic.

But Meehl also envisioned appraising a theory  $T$  through its verifiable implications under a diversity of auxiliaries  $S$ . If  $C_a^*$  and  $C_b^*$  are the corroborations currently accumulated from tests of  $T \cdot S_a$  and  $T \cdot S_b$ , respectively, with  $S_b$  appreciably different from and perhaps incompatible with  $S_a$ , is  $C_b^*$  the same as it would have been had no tests been run on  $T$  under  $S_a$ ? Or put more starkly, does cumulated corroboration of  $T \cdot S$  under one specification of  $S$  inherit any support for  $T \cdot S$  from tests of  $T$  conjoined with other auxiliaries? Surely so; but how strong should this transfer be, or more fundamentally, prior to any quantifications, by virtue of what epistemic principles should this cross-corroboration occur?

And other extensions of Meehlian  $C^*$  also need probate. Suppose that we infer  $E$  from  $T \cdot S$  and convincingly establish that  $E$  is false no matter how its *if/then* is interpreted, but find later, when reviewing  $T \cdot S$ 's implications, that  $T \cdot S$  in fact entails not  $E$  but an  $E^*$  with the same antecedent  $O_1$  as  $E$  but a modified  $O_2$  that did in fact agree with our experimental findings. Are we allowed to revise our opinion of this experiment's import to take its corroboration of  $T \cdot S$  to be what it would have been had we correctly deduced  $E^*$  from  $T \cdot S$  in the first place? (If you think not, let's hear some explanation why.) And if you do concede that, can you rationally deny that an established  $E$  generally gives post-hoc support to theories discovered or specifically developed to entail  $E$  only after  $E$  was learned?

My questions here are far from rhetorical. Nor do I suggest that they have easy answers if only we pause to ponder them. I do, however, urge that they show Popperian or even Meehlian corroboration to be massively incomplete as an account of the conditions under which our nonobservational beliefs are rational.

Of course Meehl didn't expect us to think that the 1990 design sketch of his new take on corroboration fully specified a production model. Maybe objective standards could be developed for setting interval boundaries, with formulas  $C_i$  and  $C^*$  complexified as need requires. But what would be the criterion of metatheoretic success at that? It appears, though not made clear at outset, that Meehl wanted his cumulated corroboration to be an objective measure of a theory's *verisimilitude*. This was a huge aspiration, and Meehl deserves amazed respect not just for its audacity of conception but also for undertaking many difficult pages in (1990*b*) to argue how a theory's verisimilitude should have diagnostically informative effects on  $C^*$ . But without some explication of theoretic verisimilitude sturdier than a theory's conceptual similarity to some of Omniscient Jones' verbalized knowledge, Meehl might have done better to develop his revisionist corroboration in greater operational detail before proffering it to market.

However, there is a deeper objection to Meehl's corroboration that would remain even were verisimilitude pellucid, namely, his apparent intent to decouple corroboration from practical epistemology. Why should we care how strongly a theory  $T$  has been corroborated or dis-corroborated? Some of us -- and I suspect that includes you -- want to optimize how strongly we *believe* some of the theories we contemplate, because that affects at least a little and sometimes greatly what we do, with consequences to which we are not indifferent. Had Meehl learned that his state's Mayo Clinic had chosen his  $C^*$  for summarizing results from a series of clinical trials appraising the hypothesized effectiveness of newly conceived medical treatments, wouldn't he have felt an obligation to coach these researchers on how their obtained  $C^*$  ratings should be allowed to influence adoption of these procedures? Corroboration without consequences for belief is as sterile as phone sex.

Actually, either Meehl's (1990) apparent sundering of corroboration from confirmation was unintended or he later readjusted his position on this. For in (1997) he again treated 'confirmation', 'probability', and 'corroboration' as largely equivalent at least in their qualitative concern even if not in technical detail. Cf. *inter alia* "... a question concerning the probability, confidence, confirmation, or corroboration of  $H$  ..." ( p.407), and "... better to have an index that appraises theory performance (corroboration, confirmation, degree of support) than ..." ( p.417), Most telling is his p.415 reference to "my confirmation index  $C_i$ " which earlier he had so strongly insisted was Corroboration. It remains somewhat ambiguous whether he refers by these locutions to test-induced changes in a theory's support, or to resultant (accumulated) levels thereof as these quotations prior to the last seem to favor. But that can be left for context to resolve; the salient point here is that we can now address a rather important issue of confirmation *as* an issue of confirmation, for Meehl along the rest of us, without need to double-track that in separate terms for a Meehlian (*contra* Popperian) "corroboration" whose nature is hinted at but not tenably defined by Meehl's formulas for  $C_i$  and  $C^*$ .

## Disagreements

Some mild demurrers aside, this review of Meehlian metatheory has so far been commendational. But there are large gaps in Meehl's evolving outlook, comparable to drafting the body sculpture and interior accouterments for an advanced automotive design while neglecting to allocate space for motor and fuel. The esthetic 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 methoretical 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.<sup>12</sup> 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.<sup>13</sup> 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  $K$ s are encountered singly in sequence could be viewed as concatenated corroboration from implicit repetitious testing of "Most  $K$ s 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

---

<sup>12</sup> What I find on "discovery" by word search in Meehl's published articles is mainly reference (notably 1990b, p.33; 1990c, p.137; 1992a, 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, (1992c) 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. "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.)

<sup>13</sup> 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.

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*. (For a deeper thought-experiment on inductivist vs. hypothesis-testing outlooks on inferring generalities from sample data, see footnote #16.) 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 (nevermind 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<sup>14</sup> should satisfy us unless it too has been explained, albeit neither should we foreclose the possibility of deeper explanation.<sup>15</sup> In his published work, Meehl stayed well clear of induction's justificational quicksand, but in (1990*b*), 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. 366-384. 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 *a 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.

---

<sup>14</sup> "Explanans": (a) An assertion proffered to explain something; or (b) the state of affairs so asserted.

<sup>15</sup> 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?

Actually, Meehl was involved in discovery-oriented theory development throughout his career, starting with his research on the MMPI and acknowledged in his latter-day side remark, 'I believe strongly in "exploratory" and "refined folk-observational" knowledge' (1990*c*), 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 (1990*c*, 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?

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 (1990*b*, 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.<sup>16</sup> That

---

<sup>16</sup> 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  $\varphi$  of these items' joint distribution in the unbounded population of which your observed subjects are a finite sample should be appreciably non-null. ( $\varphi$  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

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. 1990c, 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

---

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  $\varphi$  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 negligible. 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  $S_1, S_2, \dots, S_i, \dots, S_n$  of increasingly large samples drawn randomly without replacement from your database. The scores observed in each  $S_i$  yield a sample estimate  $\hat{\varphi}$  of the population  $\varphi$ -values together with an appraisal of their uncertainty, which advise you to replace  $T_i$  by a  $T_{i+1}$  positing updated  $\varphi$ -values from which you then predict  $\varphi$  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  $\varphi$  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  $\varphi$ -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  $\varphi$ '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?

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 deduced 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. 1990*c*, 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. 1990*c*, 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  $E = \text{'Either not-}D \text{ or } T\text{'}$  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.

Let us put flesh on the skeleton of this argument by exploiting the fragment of Meehl's theory of schizophrenia (his "dominant gene conjecture") mentioned above under Theme C. As briefly reviewed there, testing this foremostly requires a binary symptom of schizotypia for which, in a selected population of prospective research subjects, the true-positive and false-negative rates are known or, more realistically, well-estimated. Given this information, together with certain auxiliary assumptions about schizotypia's independence of

environmental influences and vanishingly small probability that a schizotypic child's parents are both schizotypic, Meehl drew upon established principles of Mendelian genetics to predict, for each of several categories of the blood relatives of subjects in a sample diagnosed with high confidence to be schizophrenes, and using convergent estimations of just one parameter, what proportion in each category tests positive for schotypia. I'm not sure how conclusively Meehl managed to appraise this, but let's assume that his empirical findings matched his theory's predictions well within sampling-noise tolerance. So Meehl's theory is a clear winner -- right?

Maybe not, once the news gets out that my evil twin, WR2, had proposed a competing theory of schizophrenia appealing, just like Meehl, to a defective hereditary condition but positing this to be not a bad chromosomal gene but corruption of a distinctive plasmid having a special propagation mechanism: Each cell tolerates just one plasmid of this special type, which adheres to the cell's inner membrane in a matching socket. During ordinary growth mitosis this socket with anchored plasmid duplicates with one copy going to each side of the dividing cell, whereas during meiosis this duplication is suppressed so that only one of the two resulting gametes contains this anchored plasmid. So hereditary transmission of this special plasmid is just as Mendelian as inheritance of a gene. Once WR2 further hypothesizes that a defect in this plasmid causes schizotypia and supplements that with Meehl's auxiliary assumptions, his schiz-theory predicts the same pattern of schizotypic symptoms among the relatives of confirmed schizophrenes that corroborates Meehl's version. So how do  $TsM$  (Meehl's theory of schizophrenia) and  $TsR$  (WR2's counterpart thereof) compare in their confirmation by data -- call it  $Dr$  for data-on-relations -- that affirms their mutually predicted pattern of schiz-symptom propensities among relatives of schizophrenics?

First, let's evade the issue of post-hoc predictions. No one who believes discovery to have probative value can agree that a  $T$  gets no epistemic credit from entailing a  $D$  known prior to  $T$ 's conception. But metatheory of that, which is underdeveloped apart from Bayesian confirmation which doesn't cope with conception onsets, is not feasible to probe here. So let's assume that WR2 proposed  $TsR$  before we learned  $Dr$ . Even so, might the pre-test credibility of  $TsR$  at least be discounted by our knowing that WR2 created this using Meehl's derivation of  $Dr$  from  $TsM$  as a model to modify? I intuit that this does indeed downgrade whatever outset plausibility  $TsR$  would have otherwise; but that too belongs to the metatheory of initial credibilities crying for serious discussion elsewhere. So to get on with what can be addressed here even if inconclusively, we pick up as WR2 admits that  $TsR$  was designed to give  $TsM$  some competition and that it does seem somewhat less plausible than Meehl's version, but that even so it *might* be true and merits Popperian appraisal just as much as do less audacious theories.

So how are the credibilities of  $TsM$  and  $TsR$  affected by verification of their mutual prediction  $Dr$ ? By Meehl's proposed  $C_i$ -formula they should be corroborated equally, though how that increases their respective credibilities remains unclear. But no matter -- this is the wrong question to be asking. The salient issue, when assimilating the epistemic import of a verified prediction common to competing theories, is how our opinions on

the points of disagreement among them should be *differentially* adjusted. *TsM* and *TsR* propose different hereditary mechanisms for delivering *Dr*, and although these are not strictly incompatible, the conjunction of *TsM* and *TsR* entails *Dr* only conditional on the additional premise that these two mechanisms work in perfect synchrony, i.e. that during meiosis a haploid cell gets a schizy gene if and only if it also gets a schizy plasmid. Since we surely agree that this is tantamount to impossible, the component of *TsM* (or *TsR*) that posits schizy genes (plasmids) implicitly denies *TsR*'s (*TsM*'s) component that posits schizy plasmids (genes).<sup>17</sup> So either *Dr*'s verification leaves the credibility of each theory's core posit of this hereditary transmission's nature unchanged while confirming something else in each theory, or this finding must confirm the delivery posit of one while disconfirming the other's rejection of that. We can't rationally let *Dr*-verification confirm all these schiz-delivery affirmations and denials, so which of them are winners and which are losers?? We would surely favor genes over plasmids, since chromosomal inheritance is already so well established. But given our awareness that *Dr* has confirmed *TsR* as well, should we take that also to confirm its denial of schizy genes and hence downgrade *Dr*'s corroboration of *TsM*? More broadly, how if at all should the evident conflict between *TsM* and *TsR* change our opinion of *Dr*'s epistemic import for *TsM*'s assorted components from what that would be were we unaware of *TsR* and/or other alternatives to *TsM* that entail *Dr*? If you are unsure how to answer, you have every right to be. For the bottom line of this thought-experiment is that

*We have no metatheory whatever, nor any established practice, on how to allocate rationally differentiated credibility enhancements and sometimes decrements over the assorted propositional components of a theory whose successful prediction has confirmed it as a whole.*

In particular, the hypothetico-deductive model of theory confirmation gives us no good reason to trust still-unverified predictions taken from a previously corroborated theory. We do have other reasons for cautious confidence in these, starting with the wholesale empirical induction that theories having a strong track record of past prediction successes are a good bet to continue their quality performances. But it is unprofessional to leave our understanding of nondemonstrative inference that primitive.

Accordingly, let me urge an essential startup for your proper evidential adjudication of a theory (or hypothesis, or speculation) *T*. By "evidential" I mean appraising *T* in light of some proposition *E* (possibly quite complex) inferred from *T* whose truth can be determined beyond reasonable doubt, while "infer" is whatever sequence of reasoning comprises one or more steps of form "if (that) then surely (this)" where "then surely" alludes to whatever connection between fore and aft transmits conviction for you. (I'm addressing the inferencing

---

<sup>17</sup> Here's yet another theoretic alternative: Schizy genes and schizy plasmids both exist, but never co-occur in postpartum humans because gametes receiving both never survive. Let's presume that *TsM* and *TsR* both deny this duality of schizotaxia. Were space to permit, it would be instructive to discuss this possible duality as an additional competitor to *TsM* and *TsR*.

you live by, which may or may not entirely agree with mine or with academic texts on logic, and wherein “surely” may sink to weaker grades of “probably”. And I’m putting *you* into this evaluation process because if you don’t participate at least to the extent of tracing with assent or demurrer someone else’s argument from *T* to *E* you haven’t earned the right of opinion on what in *T* would be confirmed by establishing *E*.) Finally, my subjunctive “would be” a few words back is to point out that articulating *E*’s import for *T* can and should be endeavored before incurring the possibly-considerable cost of ascertaining *E*. It may well be found that what *E* confirms in *T* isn’t altogether what you hope to learn (e.g. *E* may be sensitive mainly to *T*’s parameter *alpha* whereas parameter *beta* is your target of concern), a misdirection which, if detected in time, may be correctable by reworking your research design.

The thesis to be advanced -- one to which Meehl demonstrated implicit assent by his frequent cogent references in (1990*a,b,c*) to “derivation chains” -- is that when conjecture *T* has evidential consequence *E* (where *T* comprises all the core and auxiliary propositions invoked to reach *E*), verification of *E* directly confirms only those components of *T* that matter for deriving *E* from *T*. And ascertaining that requires you to articulate with some care the inference steps (Meehl’s “derivation chains”) by which you extract *E* from *T*. To undertake this, you should start by trying to write down all the assertions in your theory-*cum*-auxiliary-assumptions *T*, and additional propositions derived from those, that you drew upon when deriving the extant or prospective evidence *E* whose confirmation of *T* is to be appraised. (In principal, *E* should comprise all the potentially forthcoming evidence that bears on *T*; in practice, of course, *E* can be only a fragment of what is possible.) This listing is not easy, especially if you are unpracticed at it. You can’t reasonably expect to get all of what’s relevant to *E* in *T* on your first try, and neither will you manage to make explicit all the conceptual complexities that are more or less elliptically expressed by the words you do write down. But you can later fill in what you initially left out, and correct what you didn’t get right, when discovering need for that as you proceed. (If you try too hard to avoid omissions and errors at the outset, you’ll give up before getting well underway.) For simplicity, let’s call all the assertions in this initial listing “premises” of *T* even when some are consequences of others. Whenever one of these is a separable conjunction (no overarching quantifiers), split the conjuncts apart as separate premises. Next, view the set of these maximally separated premises as nodes in a directed-path diagram wherein evidence *E*, or better the set of its separable conjuncts, is also included. (We will also take the latter to be premises, as indeed they are prior to verification.) Finally, to complete Phase 1 of this analysis, put a direct path (you could draw these as a digraph, but a two-dimensional table of binary entries for node pairs is easier) from premise *i* to premise *j* just in case *i* is one of the premises, if any, from which premise *j* is derived in the inference from *T* to *E*. The best way to develop this is to ask of *E*, or of each separable component thereof, what are the listed premises directly antecedent to that in your *E*-derivation; then ask the same question for each node directly antecedent to *E*, and continue to identify direct antecedents by tracing backward in the derivation until every node’s premise is rated for whether it derives at least partially from anything else in *T*.

In Phase 2, you attempt to strip  $T$ 's components of what therein is irrelevant to their inferential production of  $E$ . Identifying what we may call the "fluff" therein is easy, once the table or digraph of direct-inference paths is in hand: A premise is fluff in your inference from  $T$  to  $E$  is just in case its node is not on any path leading to a component of  $E$ , and can be erased from your list of premises deriving  $E$  from  $T$ . Fluff should also be deleted from  $T$  unless you can adduce some additional prospective evidence which connects with it; but ignore that until you have also attempted to trim fat out of the  $T \rightarrow E$  inference. Unfortunately, the fat story proves to be more complexly obscure than could be done justice here even if, counterfactually, I were completely clear on it in my own thinking. But I'm proselytizing here not for a switch of dogmas but only for some serious interactive discussion of important metatheoretic issues more deeply than has been our superficial norm. So if I leave you perplexed, take that as provocation to join a clarification crusade.

By the "fat" in a theory  $T$ , relative to entailment  $T \rightarrow E$ , I mean the excess in components of  $T$  that qualify under Phase-1 analysis as inferentially antecedent to  $E$  but are stronger than this inference needs, such as when  $T$ , addressing processes in a group of things, presumes that all the group's members have the same value of a certain input variable at any one stage of the process whereas what follows from that for the  $E$  at issue depends only on the group means thereon. I submit that when we try to judge the confirmational import of evidence  $E$  for the diverse components of a hypothesis  $T$  that entails  $E$ , (a) only the propositions in  $T$  on which some previously uncertain component of  $E$  is directly or indirectly dependent in this derivation receive confirmation by virtue of this deductive connection (though additional consequences of the  $T$ -components so confirmed would perforce have their credibilities enhanced as well), while (b) when a component  $t$  of  $T$  not excluded as fluff can be factored as a conjunction of fat and muscle (where  $t$ 's muscle in this context is the weakest proposition entailed by  $t$  that can replace  $t$  without disrupting inference of  $E$  from  $T$ ) its fat is tantamount to fluff and earns no confirmational credit from its superfluous participation in  $T$ 's entailing  $E$ . There are loose ends to tie off before some elaboration of this proposal can be canonized, but even in its present rough outline it is worth your contemplation -- as some further comparison of competing theories  $TsM$  and  $TsR$  may show.

The parsing of  $TsM$  that enabled WR2 to produce alternative theory  $TsR$  of schizophrenia gives a nice informal example of theoretic fat in action. ("Informal" in that I'm not going to construct a confirmation-path articulation of  $TsM$  nor even try to verbalize its  $E$  precisely.) The heart of Meehl's schizy-gene hypothesis -- at least with respect to the near-relatives evidence he was pursuing (I'm dumbing down his more intricate theory somewhat to keep the thought-experiment manageable) -- is (a) that clinically diagnosable schizophrenia is rooted in some inherited feature which potentiates certain environmental conditions during maturation to bring about the clinical debility; (b) that inheritance of this feature follows the Mendelian pattern of dominant single-factor transmission through the fusion of haploid gametes that may or may not carry this factor; and (c) that the chance

of a viable human infant's having received this factor from both parents is negligible.<sup>18</sup> What follows is a .5 probability that any one child of a parent carrying the schiz-factor also receives this, from which, together with some statistical-independence presumptions about disparate schiz-factor transmissions and adjustments for estimated error rates of schizotypic symptoms, the incidence of schizotypia at various genetic distances in relatives of manifest schizophrenics can be inferred. But additional presumptions about the substantive nature of this posited schiz-factor don't matter for this deduction: *TsM*'s and *TsR*'s respective positing this to be a single gene, or plasmid, are superfluous fat laid on common muscle which presumes only that *something* in human heredity functions so and can be larded out in other imaginative ways as well.

To be sure, fat can often be converted to muscle by additional data: In the present case, testing for linkage between the schiz-factor and known chromosome markers should largely settle whether this factor has a chromosome site and hence provisionally qualifies as a gene. But just the schizotypic-symptoms-among-relatives evidence *Er* has no confirmational preference for *TsM* over *TsR*. This is not to say that, given only this evidence for the posited schiz-factor's reality, that add-on hypothesizing this factor to be a gene is no more credible than hypothesizing it to be a plasmid: What is already well-established about inheritance implies that gene is vastly more plausible than plasmid. But the test of *TsM* whose outcome is *Er* does not further increase that plausibility inequity, at least not by virtue of being a premise in *TsM* when that is corroborated as a whole by *Er*. By no means am I suggesting that Meehl should have suppressed *TsM*'s conjecture of the posited transmitter's nature or, more generically, that theoretic fat is epistemically deplorable. What I do urge is that every stage of appraising a theory *T* should include serious effort to discriminate between (a) what in *T* is well-supported, or will be if the planned research turns out confirmatory, and (b) its unsupported components that under present planning will remain so. If, for whatever reason, we feel an allegiance to some of the current fat in *T*, then our research on *T* should seek to design an evidential procedure for whose outcome this part of *T* distinctly matters. Otherwise, we should relax *T* to admit alternatives in this respect, judge whether we really care which of those is correct and, if we do, puzzle out what observations might illuminate that. Would Meehl have agreed? Surely so, at least in broad outline. For taking NHST confirmation of a statistical hypothesis *H* to be similarly strong support for the substantive *T* from which *H* professedly derives, which Meehl so strongly and repeatedly scorned, is simply one egregious but otherwise undistinguished instance of confounding a positive test result's corroboration of a theory's muscle for that prediction with corroboration of its fat.

There is a further important point to take here, a view from the edge of a *terra incognita* abyss. As I have argued in Rozeboom, 1997, p.349ff., a major facet of our real-life belief systems are the *conditionalities* therein which we verbalize by varieties of "If *p* is (or were to be) the case, then probably (to some degree generally less

---

<sup>18</sup> I'm not sure whether Meehl had any basis for presumption (c) beyond the near-negligible size of a small probability's square. But were it needed, he could have posited that a double dose of this factor is lethal to zygote development.

than certainty) so would  $q$ ". Such conditionalities among a hypothesis  $T$ 's components that hold prior to verifying an  $E$  entailed by  $T$  transduce  $E$ 's confirmation of components in  $T$ 's muscle for  $E$  into adjusted confidence in the consequents of our conditional beliefs whose antecedents are premises in this muscle. But notice how deep this drives the conundrum of warranted belief: Before working out how data on the relatives of schizophrenes might teach us more about the causes of this condition, Meehl speculated that a single dominant gene transmitted this; and if his thinking had been sufficiently astute and comprehensive (more so than could reasonably be expected in reality) he would have partitioned this conjecture as Mendelian transmission by a carrier of unspecified nature conjoined with a set of alternative possibilities for what that nature might be. And from there, his belief system should (superhumanly) have arrived at a set of conditional beliefs of form "if schizotaxia has a Mendelian hereditary transmitter, then it is likely to degree  $\gamma$  that the specific nature of this is  $\Gamma$ ", where  $\gamma$  is a strength of conditional credibility and  $\Gamma$  ranges over a set of alternatives wherein gene and plasmid are two out of many. In practice, of course, no one even as bright as Meehl could explicitly work all those out in advance. But when interpreting evidence  $E$  from relatives of schizophrenes as confirmation of the muscle in  $TsM$ ,  $TsR$ , and yet other delivery conjectures, before adjusting his prior opinions about the transmitter's nature he would by rights need to activate his latent conditional beliefs specifying alternatives for that with the same  $\gamma$ -strengths they would have had were they revived prior to learning  $E$ . Thus I put it to you: Holistic models of hypothesis testing have scarcely a clue to the subjunctive underlay of evidence interpretation. And to my knowledge, there exists no applied epistemology on the management -- acquisition and execution -- of our subjunctive conditional beliefs beyond what is implied by logical entailment. Would that this were otherwise.

There are still other issues of selective confirmation I would have liked to raise with this example, in particular (i) what we should say about an alternative to  $TsM$  that expands Meehl's pair of alleles (Normal vs. Schizy) at some determinate chromosome site into a larger plurality of alleles there that yield distinguishable grades of schizotaxia, and (ii) how the difference between a theory deploying implicitly defined (open) theoretical concepts and its corresponding Ramsey sentence would play out in this multiple alleles conjecture. But it's probably just as well at present not to heap more fuel on what I hope has already become an uneasy fire. (If I've done my job, you should be feeling some heat.) What should matter for you is the take-home message that we still have no adjudicated guidelines to refine our experientially honed intuitions on how belief changes should be allocated within a theory in response to supporting evidence. There are major metatheoretic/epistemic advances to be sought here. So if you're young, analytically skillful, and boldly independent in your thinking, consider trying to be first to succeed.

F. Meehl on cliometric theory appraisal.<sup>19</sup>

In the symphony of Meehl's distinguished career, his last movement reprises in a new key with more advanced orchestration the theme of his greatest youthful hit. In *Clinical versus Statistical Prediction* (1954b), Meehl made the case that clinical psychology can diagnose psychiatric conditions much more effectively by modern quantitative methods of multivariate estimation than by the therapist's intuitive interpretation of qualitative impressions. Then starting in the mid-'80s, initially provoked by and later collaborating with Faust (Meehl, 1984; Meehl, 1992b; Faust & Meehl, 2002), he developed the prospect of applying the same statistical procedures to diagnosing the epistemic promise of theories still wanting final adjudication, with "ensconcement" (survival longevity) taken to be the primary criterion of a theory's success. With the progenitor of cliometric metatheory (*CM*) here to give you an unadulterated overview of its nature and promise, further outsider remarks on that from me would be pointless. Even so, I would like to air one concern which I suspect will also prove bootless.

If Faust & Meehl's aspirations for *CM* achieve some fulfillment, there are both positive and negative returns to anticipate. On the upside, whenever a large assemblage of multidimensional data can be collated, modern methods of multivariate analysis can discover therein, and induce provisional explanations for, patterning far beyond what everyday perception of those event records could discern. Which is to say that much might be learned from *CM* applications, albeit just what may be harder to predict. But countervailingly, were *CM* to develop scales for predicting a theory's ensconcement, and support for research motivated by a provocative new theory were made dependent on the theory's scoring high on that forecast, work of great prospective payoff could well be stifled. Careful choice of prediction targets for *CM*'s rating scales would be crucial were they ever to be used for research evaluation in practice; and different scales with different validity criteria should be constructed for appraising different stages of progressive research. Whether a theory that has been in play for quite some time rates high on its ensconcement forecast could be an argument for continuing its support; but for new theoretic notions barely unfurled, I for one would urge that those with promise to generate intriguingly novel data configurations be encouraged for whatever abductive harvests those may yield, nevermind the odds on whether their initiating theories flourish or flounder in endurance.

---

<sup>19</sup> "Cliometrics" is the use of statistical methods to analyze historical records. I had to look it up.

Epitaph:

Meehl's autobiography communicates so much personal information that it takes some effort (at least it did for me) to appreciate how little it reveals about affect and value in his adult life beyond the satisfaction he took from intellectual achievements and occasional pique when some appeared not to achieve the professional impact they deserved. We get terse descriptions of certain stressful experiences in Meehl's youth, two life-threatening events later in life, and repeated reference to various academic colleagues as friends; but otherwise not a word about what, beyond his intellectual accomplishments, was emotionally meaningful to him. Nothing said about family later than his report of losing both parents at separate times early in his adolescence; and only from the memorium page in Meehl (1973*a*) have I learned that in midlife he also lost a wife. No mention of affect taken from music, fiction, visual arts, drama, or recreations. And not a word about the how, why, and what of his "dozen years as a Lutheran" (1989, p.374) when he had claimed his early religious indoctrination to have been negligible (1989, p.338) and later, long after discontinuing sectarian affiliation, he professed to hold no theological opinions (1989, p.378). It would be interesting to know how Meehl reconciled theological and scientific beliefs during this period.

Insomuch as Meehl manifestly prized intellectual honesty and competence above all else, it seems appropriate to close with a word on his own stature in this regard, especially insomuch as his overall brilliance may blind future historians of psychology to one important respect -- creativity -- in which he was merely good. Fortunately, since otherwise I would risk seeming disrespectful, Meehl himself adroitly acknowledged much of this in his autobiography:

At heart I am more of a knowledge-absorber, knowledge-integrator, and knowledge-transmitter than knowledge-producer. ... [big excision] ... a scientist should do what he is good at, and I am better at conceptualizing than at experimenting. My synthetic-creative talents are only somewhat superior to most psychologists (cramped by the dustbowl empiricist flavor of my Minnesota training?); but my analytic powers are, I believe, exceptionally strong and well cultivated through long association with top-caliber philosophers of science. Knowledge is advanced in several ways, and it has been my experience that there are many more psychologists who are capable of performing a clever and replicable experiment than there are high-level ideators who can create a novel concept or deeply analyze a familiar one, especially one in controversy. Living off the taxpayer, I feel it appropriate to do what I am best at, especially since (1) it's rare, and (2) I find it more fun. (1989, p.374)

Exactly!

REFERENCES

- Bolles, R. C. (1962). The difference between statistical and scientific hypotheses. *Psychological Reports*, 11, 639-645.
- Campbell, D. T. (1957). Methodological suggestions from a comparative psychology of knowledge processes. *Inquiry*, 2, 152-182.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81-105.
- Carnap, R. (1962). *Logical foundations of probability*, 2nd ed. Chicago: Univ. of Chicago Press.
- Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281-302.

- Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281-302.
- 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, S185-S196.
- Feigl, H. (1950). Existential hypotheses: Realistic versus phenomenalist interpretations. *Philosophy of Science*, 17, 35-62.
- Feigl, H. (1956). Some major issues and developments in the philosophy of science of logical empiricism. In H. Feigl & M. Scriven (Eds.) *Minnesota studies in the philosophy of science, (vol. 1)*. Minneapolis: University of Minnesota Press.
- Garner, W. R., Hake, H. W., & Eriksen, C. W. (1956). Operationism and the concept of perception. *Psychological Review*, 63, 149-169.
- Golden, R. & Meehl, P. E. (1978). Testing a single dominant gene theory without an accepted criterion variable. *Annals of Human Genetics London*, 41, 507-514.
- Hempel, C. G. (1945). Studies in the logic of confirmation. *Mind*, 54, 1-26, 97-120.
- Hull, C. L. (1943). *Principles of behavior*. New York: Appleton-Century-Crofts.
- James, W. (1897) *The will to believe*. Longmans, Green, & Co.
- Lazarsfeld, P. F., & Henry, N. H. (1968) *Latent structure analysis*. Boston: Houghton Mifflin.
- Livermore, J. M. & Meehl, P. E. (1967) The virtues of M'Naghten. *Minnesota Law Review*, 51, 789-856.
- MacCorquodale, K., & Meehl, P. E. (1948). On a distinction between hypothetical constructs and intervening variables. *Psychological Review*, 55, 95-107.
- MacCorquodale, K., & Meehl, P. E. (1954) E. C. Tolman. In W. K. Estes et.al., *Modern learning Theory* (pp. 177-266). New York: Appleton-Century-Crofts.
- Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis: University of Minnesota Press.
- Meehl, P. E. (1959) Some technical and axiological problems in the therapeutic handling of religious and valuational material. *Journal of Counseling Psychology*, 6, 255-259.
- Meehl, P. E. (1962). Schizotaxia, schizotypy, schizophrenia. *American Psychologist*, 17, 827-838.
- Meehl, P. E. (1965). Philosophy of science and Christian Theology. In J. Bodensieck (Ed.), *Encyclopedia of the Luthern Church, Vol. 3* (pp.1894-1896). Minneapolis: Augsburg Publishing House.
- Meehl, P. E. (1970). Psychology and the criminal law. *University of Richmond Law Review*, 5, 1-30.
- Meehl, P. E. (1971). Law and the fireside inductions. Some reflections of a clinical psychologist. *Journal of Social Issues*, 27, 65-100.
- Meehl, P. E. (1972). Specific genetic etiology, psychodynamics and therapeutic nihilism. *International Journal of Mental Health*, 1, 10-27. Reprinted in Meehl, *Psychodiagnosis: Selected papers* (pp. 182-199). Minneapolis: University of Minnesota Press, 1973.
- Meehl, P. E. (1973). *Psychodiagnosis: Selected papers*. Minneapolis: University of Minnesota Press, 1973.
- Meehl, P. E. (1973). MAXCOV-HITMAX: A taxonomic search method for loose genetic syndromes. In *Psychodiagnosis: Selected papers* (pp. 200-224). Minneapolis: University of Minnesota Press, 1973.
- Meehl, P. E. (1973). Why I do not attend case conferences. In *Psychodiagnosis: Selected papers* (pp. 225-302). Minneapolis: University of Minnesota Press, 1973.
- Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, 46, 806-834. Reprinted in Meehl, *Selected philosophical and methodological papers* (pp. 1-43). C. A. Anderson and K. Gunderson (Eds.) Minneapolis: University of Minnesota Press, 1991.
- Meehl, P. E. (1984). Forward. In D. Faust, *The limits of scientific reasoning*. Minneapolis: University of Minnesota Press.

- Meehl, P. E. (1989). Autobiography. In G. Lindzey (Ed.) *History of psychology in autobiography*, Vol. VIII (pp. 337-389). Stanford, CA: Stanford University Press.
- Meehl, P. E. (1990) Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, 66, 195-244. (a)
- Meehl, P. E. (1990) Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant using it. *Psychological Inquiry*, 1, 108-141, 173-180. (b)
- Meehl, P. E. (1990) *Corroboration and verisimilitude: Against Lakatos' "sheer leap of faith"* (Working Paper, MCPS-90-01). Minneapolis: Center for Philosophy of Science, University of Minnesota.
- Meehl, P. E. (1991) Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. In C. A. Anderson & K. Gunderson (Eds.), *Selected philosophical and methodological papers* (pp. 1-43). Minneapolis: University of Minnesota Press (Original work published 1978).
- Meehl, P. E. (1992) 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. (a)
- Meehl, P. E. (1992) Cliometric metatheory: The actuarial approach to empirical history-based philosophy of science. *Psychological Reports*, 71, 339-467. (b)
- Meehl, P. E. (1992) Factors and taxa, traits and types, differences of degree and differences in kind. *Journal of Personality*, 60, 117-174. (c)
- 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
- Meehl, P. E. Cliometric metatheory II: Criteria scientists use in theory appraisal and why it is rational to do so. *Psychological Reports*, 91, 339-404.
- Meehl, P. E. (2004 online) The power of quantitative thinking. Currently available at website [www.tc.umn.edu/~pemeehl/](http://www.tc.umn.edu/~pemeehl/)
- Popper, K. R. (1959). *The logic of scientific discovery*, 2nd ed. London: Hutchinson & Co.
- Popper, K. R. (1972). *Objective knowledge*. Oxford: Clarendon Press.
- Rogers, A. I., & Hoel, D. (1997). Peptic ulcer disease: Tracing science's journey through the gut. *Postgraduate Medicine online*, 102(5).
- Rozeboom, W. W. (1960). A note on Carnap's meaning criterion. *Philosophical Studies*, 11, 33-38.
- Rozeboom, W. W. (1962). Of selection operators and semanticists. *Philosophy of Science*, 31, 282-285.
- Rozeboom, W. W. (1970). The art of metascience; or, what should a psychological theory be? In J. R. Royce (Ed.), *Toward unification in psychology*. Toronto: Toronto University Press. (a)
- Rozeboom, W. W. (1970). The crisis in philosophical semantics. In M. Radner & S. Winocur (Eds.), *Minnesota studies in the philosophy of science (Vol. IV)*. Minneapolis: University of Minnesota Press. (b)
- Rozeboom, W. W. (1972). Problems in the psycho-philosophy of knowledge. In J. R. Royce & W. W. Rozeboom (Eds.), *The psychology of knowing*. New York: Gordon & Breach.
- Rozeboom, W. W. (1984). Dispositions do explain; or, picking up the pieces after Hurricane Walter. *Annals of Theoretical Psychology*, 1, 205-223. New York: Plenum.
- Rozeboom, W. W. (1997). Good science is abductive, not hypothetico-deductive. In: Harlow, L. L. & Mulaik, S. A. (Eds.) *What if there were no significance tests*. Mahwah, NJ: Lawrence Erlbaum Assoc.
- Tolman, E. C. (1959). Principles of purposive behavior. In: Koch, S. (ed.), *Psychology: A study of a science, Vol. 2*. New York: McGraw-Hill.
- Waller, N. F. & Meehl, P. E. (1998). Multivariate taxometric procedures: Distinguishing types from continua. Newbury Park, CA: Sage.

\*\*\*\*\* Appendix cut from published version \*\*\*\*\*

## Semantic Imperfections

.Whatever reservations one may have about Meehl's "verisimilitude" conception of theoretic veracity, it has the considerable virtue of seriously attempting to recognize in some detail that this is rather more intricate than a simplistic true or false. However, it does not much stretch the classical binary conception of truth to grade the partial truth of a set or conjunction *S* of sentences by some function of a binary-truth distribution over *S*. (*S*'s proportion of true constituents would be simplest, but we could also define weighted proportions if we can quantify the relative importance of *S*'s assorted constituents.) A deeper challenge is to understand how language usefully represents the external world for us with concepts that are unclear. It is an ineluctable inductive inference from close analysis of the words we live by that their meanings are always blurred at their edges and often most of the way to center, which is to say that even our simplest sentences, whether conjoined in scientific conjectures or uttered in everyday living, have some degree of truth deficit due to their imperfect clarity. Despite the ubiquity of this commonplace semantic imperfection, I have seen recent evidence that a considerable proportion of psychologists in at least some sectors of quantitative applications profess to find incomprehensible the very thought of approximate truth, which is surely the best we can expect of propositions whose meanings are murky. For how can a statement be impeccably true if some of its most salient concepts are vague? No recognized epistemic principles allow that, at least not for vagary to be pervasive in beliefs secure enough to qualify as hard-core knowledge. Yet for the most part, occasional legal conflicts notwithstanding, we live comfortably with words with flex in their meanings, and it would be nice to have some philosophically articulate understanding of how this works for us. I can only hint at that here (beyond which I could advance only a little even were space no problem), but would like to offer some start-up considerations that invite furtherance.

### (A) The Ubiquity of Vagueness and its Ilk

To begin, a few examples will remind you how universal are semantic suboptimalities in the language we live.. Consider how you conceive of what is most immediately foreground in your own life, starting with aspects of your physical body. Are you sure of just what that comprises and, if not, do you think that when you say or think "my body" (henceforth "*mb*" here) this concept nevertheless picks out a precise referent? Are your fingernails, or the dirt under them, or your body hair, or your workout perspiration part of *mb*? And if so for some portions thereof that are subsequently detached by clipping/cleaning/cutting/wiping, are those still parts of *mb*? If not, at what precise stage of the removal process do they become not so? (Mere separation doesn't suffice, at least not in my personal understanding of English: If my car, or dishwasher, or computer is partly disassembled for cleaning or repairs, I still consider the detachments as part of my car, etc., prior to reassembly albeit that raises the question, if a defective part is replaced, when during disposal does the old part lose that status?) And does *mb* include everything within the volume of space enclosed by *mb*'s skin, in particular everything in *mb*'s digestive tract? (Some or all of that can be excluded on grounds that, topographically, it is in the hole of a torus; but what say you if you think that closing the aperture at each end changes the topography?) And if an accident or surgical contrivance has left a bit of metal embedded in *mb*'s interior tissue, is this perforce a part of *mb* or might that depend on how long it has been there and the circumstances of its implantation? These examples may well seem trivial to you, and indeed they should if

you mistake me to be protesting them. Quite the opposite: Vagaries like these are to our use of language as lubricants are to the precision machinery whose bearings they keep from seizure. What we want is not to expunge them but to better understand their semantics (theory) and management (practice).

Ambiguity and vagary are just as rife in the predicates we ascribe to things in hypotheses and data statements. I'll give only one example here, but volunteer to exhibit the fuzz in any other established English predicate on which you challenge me if you genuinely can't see this yourself without help. Were you to comply with request for a description of *mb*, you would surely need little prompting to appreciate how imprecise is your identification of *mb*'s skin color, flabbiness, hairstyle, facial symmetry, slouch and their ilk. So let's consider a feature that we standardly measure, say Height. Informally, we describe the alternatives by 'tall', 'medium', 'short', which we can refine with 'extremely', 'very', 'somewhat', 'mildly', 'sort-of', and other tweak-adjectives. But when wanting to grade height alternatives more finely than this, we now know how to quantify those in terms of an elective measurement unit. Not wanting to belabor this, I shall assume that you have some sense of what it means to say that two points in space at a given time are  $r$  centimeters apart, where your  $r$ -concept professedly names a real number. (There's a lot of theory embedded therein, but that's true of all our verbalizable concepts.) And let's agree that a person's height-in-cm. is the straight distance in centimeters between two points on the periphery of that person. But which two points are those? Can we define them in a way that unambiguously identifies their locations on *mb* and anyone else we consider to have height at the times he/she/it has it? For example, do the places at which a measuring setup such as a physician's height slide touch *mb* suffice to identify these points? Not so for several reasons. First, if the contact points are at the outer edges of sole callus and head hair, we may not want to include the span of those body parts in *mb*'s height. Secondly, there will generally be many simultaneous contact points and we need to define which pair to pick. (Can do, but that component of ambiguity persists until then.) Most importantly, measuring *mb*'s height has posture requirements: Spine must be vertically straight so far as feasible; no tip-toeing or arm raising; and head angle, unless standardized in some way I suspect has never been discussed much less agreed, puts additional chanciness into what distance between which contact points is selected. What these considerations show is that *mb*'s height is actually a disposition to yield one or another length measurement depending on circumstances; and the concept's ambiguity lies foremostly in our failure to say precisely what the conditions are under which the distance between selected points on someone's body surface is that person's height. (And you're right to think that this uncertainty is utterly inconsequential under most circumstances. Yet consider the heartache and lawsuit that might ensue from failing by a barely discernable margin the minimum height requirement for a job or privilege sought by a strongly motivated applicant.)

But, you may challenge, don't at least some of our predicates, for example those of form 'the distance between \_\_\_ and \_\_\_ is  $d$ -units', wherein ' $d$ ' and '*units*' respectively identify a real number and unit of measurement, designate their target attributes exactly even if most distances (e.g. ones corresponding to an irrational number of scale units) can be conceived only approximately? I can't dismiss this possibility generically; but if someone can give convincing examples, distance measurements won't be among them. Details on this are impractical here, but note these three points: First, the two arguments of a distance-between predication are primarily *points* in space-time (though we can generalize to spatially extended objects by stipulating, say, that the distance between two wholes is their nearest-points separation); and we are never (or scarcely ever?) able to identify either of these points or point-aggregates exactly when predicating a distance-between of them. So even if some distance predicates have precise designata in spacetime reality, we can never (or scarcely ever?) make exactly true assertions with them. Moreover, our

notion of distances among points in space/time proves under analysis to be deeply theoretical even commonsensically, not to mention modern physics whose more advanced conceptions thereof don't translate into folkspeak very well. Either way, counting "units" of distance -- inches, meters, miles, etc. -- presupposes that we have a concept of what it is for two points in space-time to have exactly the same separation as two other points, as well as awareness of some specific pair of points whose separation is the standard for such comparisons. I venture that you will have trouble persuading yourself that either of these notions is completely clear to you.

(B) Conceptual inexactitudes

Ambiguity, vagueness and imprecision, the highest grades of conceptual imperfection before that slumps into obscurity and muddle, come in many variants that overlap and often co-occur. And to this list of familiar conceptual fuzzies, I shall append another of considerable importance but not heretofore adequately recognized, namely, multiple reference. To distinguish among these, consider some examples:

[a] My weight [at this very moment] is exactly 210.72864 lbs.

[b] My present weight is roughly (approximately, nearly, almost) 211 lbs.

[c] My present weight is between 210.0 lbs. and 211.0 lbs.

[d] Last month it was 217.

[e] My present weight is exactly five pounds less than my brother's.

Once we agree on a referent for the 1<sup>st</sup>-person pronoun therein (how that is fixed is irrelevant), assertion [a], though untrue, is about as precise as we ever get in everyday life. But if its square-bracketed qualifier (thereafter simply 'present') were omitted, (a) would be either elliptic or vague (actually, both) according to how tight a temporal window its present tense takes from context. (Also, its numeric precision cannot expunge the modest inexactitude in 'lb.'). In contrast, [b] is vague under all the listed choices for its adverb which, however, differ in degree of vagueness: "Roughly" is more diffuse than "nearly" while "almost" is even less vague than "nearly", in part because it is rather precisely one-sided -- declaring something to be almost someway generally implies that it falls a little short. In contrast, [c] is no more vague or ambiguous than [a] but is less precise, showing that although vagaries are imprecise, imprecisions needn't be vague.

Though not as idiomatic as might be preferred, [d] illustrates several facets of conceptual obscurity. Most obvious is its omission of a measurement unit and the attribute being quantified. When context makes near-certain that 'it' in this utterance abbreviates, say, 'my weight in pounds', [d] is elliptic. Otherwise, it is vague and/or ambiguous according to the alternative clarifications suggested by context. If we have been discussing my dietary purchases and weight-control workouts, you might interpret the numeral in [d] as ambiguating over dollars spent, weight in pounds, and miles on treadmill. And were I to indicate that I meant dietary purchase, that would still leave vagueness in this category's definition, and imprecision in how my

\$217 was distributed over what items of this kind. [*d*] is also vague in its temporal-location reference unless context supplies precise chronometric boundaries for that and, if those pick the entire month or an appreciable subinterval thereof, how the single number 217 which, say, I now clarify is weight in pounds is extracted from my weight trajectory over that period. Were I to clarify by reporting that 217 is my average weight in pounds over a morning and evening weighing each day, this vagueness would shrink to a modest ambiguity that finally reduces to maximally feasible precision if I further pass on what technical version of "average" I used and how it was calculated.

Finally, [*e*] illustrates a genus of semantic suboptimality still not adequately comprehended even by technical philosophy of language, though Russell's theory of definite descriptions made it a serious analytic challenge which rightly continues to attract philosophic attention. Grammatically, 'my brother's' in [*e*] is elliptic for 'my brother's present weight in lbs.', which is a nominal whose job is to designate a unique attribute of a unique object designated by 'my brother at present'. Presuming 'brother of' to be ideally precise (that it's not doesn't matter here), it is entirely possible, even if unlikely, that [*e*] is true at least as commonsense understands "true" so long as I do in fact have just one brother. But what if, unknown to me, I actually have no brother (he was incinerated just hours ago in a horrendous accident), or more than one (I've never learned of my adolescent parents' pre-marriage son given up for adoption)? There seems to be general philosophy-of-language agreement that [*e*] cannot be true absent a corporeal brother, though whether lack thereof makes [*e*] semantically ill-formed or well-formedly false is still controversial. (I have argued for the latter in Rozeboom, 1972.) More challenging is to clarify [*e*]'s semantic status when several things -- call the set of these '**Bs**' -- would each fully qualify as the referent of 'my brother at present' -- abbreviate this expression as '**Bmp**' -- were it not for the others, equally qualified, in **Bs**. Several alternative possibilities for **Bmp**'s semantic relation to **Bs** merit consideration:

[*i*] Since a designating expression's referent must be unique, **Bmp** perforce has no referent in this case unless this concept can be augmented by something that eliminates all but one of its otherwise qualified referents; and until then [*e*] has the same truth-status as it would have were its defining predicate to have no instances.

[*ii*] **Bmp** may signify **Bs** as name of the class, or alternatively, shattering our classic uniqueness-of-reference posit, may simultaneously designate each object in **Bs**.

[*iii*] Some random process might select one winner from the set of entities qualified to be **Bmp**'s referent. This could be a new pick each time its context of usage changes, comparable to our use of demonstratives, or fixation of an initial random pick as proposed by at least one major-league logician in recent times (cf. Rozeboom, 1962, for reference).

[iv] Contrary to traditional philosophy-of-language presuppositions, intentional-aboutness relations may in general be many-many and graded. In particular, perhaps a given concept can have multiple referents under varied degrees or qualities of context-sensitive representational connection. Multiple reference must not be confused with ambiguity. The latter is a complication of word/meaning evocations; multiple reference is complexity in how meanings may relate to what they are about.

Prospect [i], philosophic orthodoxy, is far too simplistic to capture the nuances of semantic reality; why [ii] doesn't work I'll leave as a easy exercise for you though its second version has some merit as an idealized precursor of [iv]. And [iii] is too magical to take seriously, though the comparison to demonstratives merits discussion before dismissal. But [iv] is a framework within which a realistic theory of semantics has room to grow, and has already been put into play by the open-concept/implicit-definition construal of theoretical concepts accepted by Meehl and most(?) philosophers of science today. According to this outlook, if I have several brothers but just one whose weight exceeds mine by exactly five lbs., then that one is the referent of **Bmp** in [e]. Whether more generally it might designate just the one whose weight exceeds mine closest to the stated difference, or multiply designate each of my brothers to a degree commensurate with our weight-difference's proximity to the contrast claimed, merits thoughtful semantic-theoretic contemplation. Such unshackling of reference should also yield graded qualities of proximal truth albeit, like Moses in the wilderness, I can only lead toward that promised land without expectation that I myself will ever help to domesticate it. (I do, however, have some salient route maps that I would proffer were circumstances to permit.)

{{ A once-planned map sketch for the journey may still be added here should any expressed interest arise. }}

My attempt here to clarify by example the differences among traditionally recognized types of semantic imprecision would be poor show if I did not also try to abstract their respective natures. So here goes: In first approximation, a statement *S* is "ambiguous" in some linguistic context if it brings several verbalizable interpretations (different meanings) to mind while context neither picks just one as intended (which might reclassify *S*'s occurrences as elliptic here rather than ambiguous) nor urges that *S* be construed as their disjunction. In contrast, *S* is "vague" in contexts wherein it conveys a haze of indistinctly many propositional alternatives, more than can be explicitly verbalized but similar to one another in some intuitive way, so that *S*'s meaning in such instances seems best explicated as a weighted disjunction over these. (I am far from convinced that this attempt to distinguish "ambiguous" from "vague" is fully adequate, but at least it's a start.) As for imprecision, this is not just semantic blur but is also relative to linguistic intent. My

characterizing the interval predication in [c] above as imprecise presumed a context of usage calling for point information. But it would be as precise as [a] were a whole-number interval to be wanted. Actually, it is surprisingly difficult to define 'precise' precisely. In essence, a sentence  $S_a$  is more precise than sentence  $S_b$  with respect to a topic/theme/aspect/concern/query  $Q$  just in case every sentence entailed by  $S_b$  and just about  $Q$  is also entailed by  $S_a$  but not conversely. But I'm not able to make precise my intuitions on what it is for a sentence to be *just* about some  $Q$ . (Aboutness I can corral, but tidy boundaries for  $Q$  elude me.)

Although hard sciences generally treat semantic inexactitudes as cognitive inferiorities to be minimized so far as feasible, the four varieties just overviewed differ considerably in the concept-management demands they place on us. In general, an imprecision range is as easy to specify as are point values within the range -- which is to say not that these are trouble-free, but only that their semantic problems/imperfections are of the same sort and degrees of severity as points. (Actually, specifying a multidimensional range is harder than just naming a few boundary points, but usually not much when a multidimensional enclosure is specified by parameters in a few well-behaved algebraic functions.) And ambiguities are routine to resolve in research areas whose participants are willing to think carefully about what they are trying to say, albeit our discipline could well profit from some training in concept analysis. (It's a pity that Meehl, who was superb at this, was never commissioned to work up educational procedures that could teach students to appreciate and maybe even practice this important intellectual skill.)

Vagaries, on the other hand, tend to work themselves out as need arises. But tightening loose concepts generally comes at a price in establishing consensual refinements and contriving to detect and report those accurately. (Imagine the joy and compliance that would ensue if government regulations were to mandate that Height in physicians' medical records be standardized on its finer indeterminacies cited above.) So in general, exactitude should not be obsessed much beyond current needs, practical or theoretical, albeit with different locutions provided for different levels of wanted exactitude such as already carried to a high art in our technical language for quantitative measurement. And as technical discriminations become honed, commonsense vagaries evolve into the elective imprecision of precisely bounded intervals. (You knew that already, of course; but I should say it for the record.)

As for multiple reference by open concepts, this is not an avoidable suboptimality in our use of language but an inherent consequence of how the world and our knowledge of it works. Whatever we know only through its discernable effects on our sensory surround cannot be cognitively distinguished from other things that have those same effects until such time as we detect and conceptually assimilate additional effects that differentiate them. But there is an important caution to take from this: When developing explanatory theories, we should be frugal in attributing to our theory's posited source entities more attributes than make a difference for our current evidence supporting them. Excessive fat can be as debilitating to a theory's epistemic health as it is to a theorist's physical wellbeing