Meehl on Theory Appraisal

Ronald C. Serlin
University of Wisconsin, Madison

Daniel K. Lapsley
University of Notre Dame

Meehl provides a thought-provoking extension of his seminal work on the hazards of null-hypothesis testing (Meehl, 1967) and the difficulties of detecting cumulative progress in psychological research (Meehl, 1978). In part, his article is intended as a response to our earlier article (Serlin & Lapsley, 1985) in which we attempted to account for slow progress within psychology and also the problem inherent in testing a null hypothesis that is always false. We dealt with the problem of slow progress by an appeal to the Lakatosian reconstruction of science. We attempted to resolve the hypothesis-testing problem by proposing a "good-enough principle," which has the effect of stiffening the observational hurdle that a theory must overcome in order for an experiment to provide corroboration for a theory under test. By specifying a good-enough region, one is able to perform a statistical test of a hypothesis that is not always false and, at the same time, to satisfy Popper's requirement regarding what is to be accepted as factual.

Although much important ground is covered, two main points seem to emerge from Meehl's article. First, he outlines two criteria the satisfaction of which would seem to justify a rational, Lakatosian defense of a theory (the "strategic retreat"). Meehl calls these criteria the Lakatos principle and the Salmon principle ("damn strange coincidences"). Second, Meehl wants to use the language of good enough in the context of theory appraisal. A theory is corroborated, according to Meehl, if numerical predictions are "close enough," and he provides a corroboration index, absent any obvious appeal to significance testing, to estimate when a theory is corroborated by empirical data.

There is much to admire in this article. Unfortunately, given the limitations of this forum, we must restrict our commentary to those features that, in our estimation, could bear another look. Although we have attempted to use Lakatosian formulations to account for growth and progress in psychological science *and* to fortify the rationality of theory appraisal using significance testing, Meehl invokes the spirit of Lakatos only to deal with the problem of theory appraisal. This leads to two problems. The failure to provide a sufficiently rigorous Lakatosian account of scientific growth ultimately undermines any attempt to provide an alternative methodology of theory appraisal. Indeed, as Lakatos (1978) pointed out, "Theories cannot be appraised without a theory of scientific growth" (p. 159). In addition, this same failure to incorporate growth in theory appraisal weakens Meehl's appeal to Bayesian statistics in attempting to "numerify" the rationality of strategic retreats. We take up each of these issues in turn.

The concept of growth is critical to "Popperian" accounts of scientific rationality. Indeed, under this view, growth is the defining characteristic of science. As Lakatos (1978) noted, "it is the progressing problematic frontiers of knowledge, and not its relatively solid core, which gives science its scientific character" (p. 174). And the characteristics of growing science are excess content, rather than content, and excess corroboration, rather than corroboration (Lakatos, 1978). The Lakatosian distinctions concerning the "acceptability" of a theory are enormously helpful in illustrating this point (see Lakatos, 1978).

Acceptability$_1$ refers to "boldness," or excess empirical content. A theory, once proposed, is initially appraised in terms of its boldness. A bold theory specifies novel potential falsifiers or has excess empirical content over a theory it challenges. If this obtains, scientists accept$_1$ the theory into the body of science. The key point, however, is that "clearly one cannot decide whether a theory is bold by examining the theory in isolation, but only by examining it in its historico-methodological context, against the background of its available rivals" (Lakatos, 1978, p. 171).

Bold theories (accepted$_1$) must next undergo severe tests. The severity of a test is also a comparative matter. Given two theories, T_1 and T_2, a severe test of T_1 (relative to T_2) tests the excess content of T_1 over T_2. A theory is corroborated (relative to T_2) if its excess content is corroborated. Hence, "severity and corroboration are binary relations between the tested theory and some touchstone theory" (Lakatos, 1978, p. 183). Although a theory may be accepted$_1$ if it has excess content over a rival, a theory is *accepted*$_2$ if it has excess corroboration. This makes clear that, for Lakatos, scientific rationality depends on problem shifts and growth, and this hinges on comparative-historical appraisals of rival theories. According to Lakatos:

> One of the most important features of the two methodological appraisals of theories is their *historical* character. They depend on the state of background knowledge: the prior appraisal on the background knowledge at the time of the proposal of the theory and the posterior appraisal also on the background knowledge at the time of each test. (p. 178)

It is the theories and the growth of knowledge that they produce that are appraised conjointly, rather than the theories "in light of the evidence" per se. Consequently, it is wrong, in Lakatos's view, to think that the verisimilitude of a theory (in light of the evidence) must be judged in isolation of historical considerations. It is a "deeply entrenched dogma of the logic of justificationism that evidential support depends on the theory and the evidence and not on the growth that they represent in relation to former knowledge" (Lakatos, 1978, p. 183). Hence, notions of evidential support and corroboration are always judged in light of historical comparisons with rival theories.

This brings us to *acceptability*$_3$. Lakatos suggests that acceptability$_1$ and acceptability$_2$ adequately capture the Pop-

perian logic of scientific discovery. Nonetheless, acceptability$_3$ refers to the future performance of a theory, its measure of evidential support, reliability, or trustworthiness. A theory is accepted$_3$ if it is judged to yield reliable predictions. Hence, the reliability and consistency of predictions determine the acceptability$_3$ of theories, and, intuitively, there is some sense that the more acceptable$_3$ a theory is, the greater is its verisimilitude (Lakatos, 1978). Our hunch is that Meehl is mostly concerned with this feature of theoretical acceptability (see, e.g., his discussion of "consistency tests" and also the *Spielraum* index).

However, Lakatos (1978) noted two serious shortcomings with acceptability$_3$ and its claim on verisimilitude. The first is that acceptability$_3$ gives us *very limited guidance* in choosing among theories in the body of "most reliable theories," all of which have stood up to severe tests. This is so because one can judge the reliability or verisimilitude only of eliminated theories. These are appraised in light of the present theories, which are the ultimate standards of the moment. And because corroboration is adjudged comparatively in terms of predecessor (or superceding) theories, one cannot devise any metric of "degree of corroboration" for the body of present "most reliable theories." Consequently, for these theories, "there is not and cannot be any 'degree of corroboration'—indeed the expression 'degree of corroboration', in so far as it suggests the existence of such a metric, is misleading. . . . But where corroborations of two theories are incomparable, so are their reliabilities" (Lakatos, 1978, p. 185).

The second shortcoming of acceptability$_3$ is that it is *unreliable*. Lakatos noted that even when comparisons are possible, one can easily conceive of conditions which would make the estimate of verisimilitude by corroboration false. In addition, "the success of scientific theories may be such that each increase of truth content could be coupled with large increases in hidden falsity content, so that the growth of science would be characterized by increasing corroboration and decreasing verisimilitude" (Lakatos, 1978, p. 185).

With this description of the three acceptabilities, we are now in a position to evaluate certain of Meehl's claims. First, consider Meehl's Lakatos and Salmon principles. The Lakatos principle says that we are warranted in continuing to conjecture that our theory has high verisimilitude when it has accumulated "money in the bank" (i.e., by passing severe tests and, accordingly, by accumulating a "good track record"). The Salmon principle says what a good track record amounts to—it is one where a theory makes successful, close-enough, near-miss predictions of events that, absent the theory, would have low prior probability. It should be clear that the two Meehlian principles describe Lakatos's acceptability$_2$. However, Meehl provides no grounds for accepting bold, new theories that have no track record or "money in the bank" into the body of science, that is, no grounds for appraising conjectural knowledge. Completely absent is any notion of acceptability$_1$. But, as Lakatos pointed out, scientific rationality also allows one to embrace a theory even though there is not a shred of evidence in its favor, as long as this prior appraisal reveals excess content.

At first blush this might seem like only a "friendly amendment" to Meehl's argument. But the problem with the Meehlian principles goes deeper—they lack the comparative, historical dimension that is so crucial to the Lakatosian account of scientific rationality. Note, for example, what counts as a bold theory for Meehl: a theory that predicts facts that, absent the theory, would have low prior probability. But boldness for Lakatos is not this, but *excess* content vis-à-vis a rival, touchstone theory. "A theory which has no more potential falsifiers than its background theory has at most zero 'excess falsifiability' " (Lakatos, 1978, p. 171). In this context predicting "damn strange coincidences" may not be decisive if considered in isolation from rivals. A theory that entails some facts that have low prior probability absent the theory is not also necessarily one that has excess empirical content relative to a touchstone theory. In other words, a theory that satisfies the Salmon principle may not satisfy the criteria for acceptability$_1$. Further, the crucial Lakatosian point must be emphasized: "The only admissable positive evidence for a theory are the corpses of its rivals" (Lakatos, 1978, p. 184)—and this no matter how well a Meehlian theory predicts an unlikely event.

The historicocomparative argument also undermines Meehl's notion of "money in the bank" and what is to count as a severe test. On Lakatosian grounds, it is not "money in the bank" or a track record that is decisive, but excess corroboration over rivals. Note also Meehl's notion of "severe test." For Meehl this amounts to a risky theory that passes "consistency tests." However, a severe test of T_1 is always relative to the touchstone T_2. A theory is corroborated if its excess content over T_2 is corroborated. A theory that yields reliable and consistent predictions need not be also one that yields excess corroboration. Parenthetically, it is not Popperian to tell a scientist to "aim at highly reliable theories," insofar as this dictum ignores the requirements for scientific growth.

Meehl's attempt to estimate verisimilitude by means of a corroboration index is also problematic, for many of the reasons already noted. Such an index provides very little guidance in choosing among the most reliable theories, and it is unreliable. It is not perverse to think that a more corroborated theory can have less verisimilitude. As Lakatos (1978) noted, "Precise, numerical estimates of degrees of 'reliability' are so unreliable as to make any such estimates utopian; moreover, even non-numerical formal expressions are misleading if they suggest that they may lead to general comparisons of any real value" (p. 193).

The Spielraum corroboration index seems particularly utopian to us. This index seems to have been motivated by the apparent appeal of Bayesian statistics as an alternative to traditional significance testing. Meehl has a particular abhorrence to the "weak" use of statistics, wherein the theoretical values, "rather than being positively generated by an affirmative substantive theory," are instead specified only by the null hypothesis; this stands in contrast to the "strong use of a significance test," where one tests "whether the distribution of observations is compatible with the predictions of a substantive theory." But it appears that Meehl deploys his examples (e.g., the cholera example) only to introduce the notion of Spielraum, for later he writes:

> It is crucial in my argument that this low tolerance is not best judged by traditional significance testing, whether of the strong or weak kind . . . It would be unfortunate if accepting some form of the good-enough principle that still emphasizes significance

> testing, especially of the weak kind, . . . should blunt the attack on that tradition.

So what, indeed, is Meehl doing?

Meehl cannot merely be using his technique for the purposes of estimation, because a uniform prior over a finite range yields the same posterior estimates as does traditional maximum likelihood estimation (Kendall, 1948, p. 179) and the same Bayesian credible interval as does the traditional confidence interval (Phillips, 1973, p. 259). So let us assume, as he claims, that he is "attacking the whole tradition of null-hypothesis refutation as a way of appraising theories" and that his Bayesian reasoning provides an alternative testing methodology. Meehl feels that although his method resembles the traditional flabby significance test, it is actually much stronger than that, because it asks how likely it would be by chance that the correlation would be picked out of the a priori interval. But can this method qualify as a strong use of statistics? There are many reasons why it cannot.

First, Meehl's continual appeal to the principle of indifference falls prey to his own criticism that the (a priori) theoretical values are merely being supplied by the null hypothesis. Secondly, Bayesian statisticians are not as sanguine as Meehl about the principle of indifference. For example, de Finetti (1974) stated, "Bayesian techniques, more or less developed into imposing mathematical machinery, are often applied as such, using standardized 'models' and standardized 'prior distributions,' instead of carefully keeping realistic adherence to the specific features of each particular case. . . . The choice must express our true opinion, . . . specifying the case and the reason" (p. 117). Bayesian statistics allow us to change our beliefs in the face of evidence, so that the choice of prior distribution should reflect the state-of-the-art of the science. However, the Meehlian use of the principle of indifference reflects an ahistorical (and hence non-Lakatosian) aspect that we found wanting in our comments on the Lakatos and Salmon principles. Because the data in the example cannot be regarded as arising from a uniform distribution, to test the cholera hypothesis (for example) on the basis of a uniform prior would only allow the conclusion that the underlying mechanism is nonaccidental (see Good, 1969). As Good noted, "As always in the Bayesian testing of a hypothesis we must make some formulation of the rival or non-null hypothesis, besides expressing the null hypothesis itself with some precision" (p. 30). Hence, even in the area of hypothesis testing, the Meehlian methodology is seen to lack the necessary historical comparisons with touchstone theories.

These difficulties notwithstanding, let us take Meehl at his word that he is providing a quantitative index of corroboration that is free of any feature of significance testing. Is this index sufficient? We think not. Meehl's own argument against using a measure of effect size for theory appraisal, and Chow's (1988) arguments that effect-size measures are not a satisfactory alternative to the significance test, can equally well be applied to Meehl's index. First, Meehl notes that the effect size could err on either the high side or the low. So, too, can his corroboration index, because it is based on sample values. Second, in none of Meehl's examples does he include the crud factor; yet, as he notes, "what we need to know, in appraising our theory, is how the correlation stands in relationship to the crud factor . . ." Third, as shown earlier to be true of all corroboration indices, Meehl's index provides little guidance in choosing among theories (this point is illustrated shortly). Fourth, as Chow (1988) asserted in the context of criticizing effect-size measures, "the issues should be (a) whether the criterion is well-defined and (b) whether the criterion would mislead its users. It can be argued that the use of the significance test is more satisfactory with regard to the latter issue" (p. 109).

Let us reanalyze Meehl's cholera example in light of the crud factor. Now the epidemiologist's prediction is that $r_{xy \cdot z}$ should fall in the crud-factor range $(-.3, .3)$, which leads directly from partial-correlation algebra to the prediction that r_{xy} should fall in the range $(.44, .65)$. When the observed correlation falls in this range, we have a strange coincidence to the extent of $p < .30$. Next, let us say that the other epidemiologist, pursuing the notion that some aspect of poverty (such as proximity to the canal, leading to bites from cholera-bearing rats or mosquitoes) leads to cholera, feels that when cholera incidence is statistically removed, the resulting partial correlation between poverty and canal water consumption will be on the level of the crud factor. Then the same crud range for $r_{xz \cdot y}$ leads to r_{xy} falling in the range $(.54, .76)$ and again a coincidence $p < .30$ when the observed correlation obtains. We are hard pressed, on the basis of the evidence provided by the corroboration index, to choose between the two theories.

What is needed is both an index and a corresponding significance test, and we feel that we have indicated this methodology (Serlin & Lapsley, 1985). Indeed, Meehl's sustained attack on traditional significance testing (rather than considering the possibility that significance testing can be fortified with the good-enough principle) recalls a good point raised by Kempthorne (1971, p. 489), with which we end our commentary:

> There are vast obscurities in the whole matter, but these are not resolved by converting the procedure into another which has superficial resemblance. Nor are they resolved by pointing to obvious misuses. Nor are they resolved by setting up the straw man, that users of tests of significance regard them as a universal panacea.

Note

Ronald C. Serlin, Department of Educational Psychology, School of Education, University of Wisconsin, 1025 West Johnson Street, Madison, WI 53706.

References

Chow, S. L. (1988). Significance test or effect size? *Psychological Bulletin, 103,* 105–110.

de Finetti, B. (1974). Bayesianism: Its unifying role for both the foundations and applications of statistics. *International Statistical Review, 42,* 117–130.

Good, I. J. (1969). A subjective evaluation of Bode's law and an "objective" test for approximate numerical rationality. *Journal of the American Statistical Association, 64,* 23–49.

Kempthorne, O. (1971). Response to critiques. In V. P. Godambe & D. A. Sprott (Eds.), *The foundations of statistical inference* (pp. 470–492). Toronto: Holt, Rinehart & Winston.

Kendall, M. G. (1948). *The advanced theory of statistics.* London: Griffin.

Lakatos, I. (1978). Changes in the problem of inductive logic. In J. Worrall & G. Currie (Eds.), *Mathematics, science, and epistemology: Imre Lakatos philosophical papers* (Vol. 2, pp. 128–210). Cambridge, England: Cambridge University Press.

Meehl, P. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34,* 103–115.

Meehl, P. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46,* 806–834.

Phillips, L. D. (1973). *Bayesian statistics for the social sciences.* New York: Crowell.

Serlin, R. C., & Lapsley, D. K. (1985). Rationality in psychological research: The good-enough principle. *American Psychologist, 40,* 73–83.