Rationality in Psychological Research

The Good-Enough Principle

Ronald C. Serlin *University of Wisconsin—Madison*
Daniel K. Lapsley *University of Notre Dame*

ABSTRACT: This article reexamines a number of methodological and procedural issues raised by Meehl (1967, 1978) that seem to question the rationality of psychological inquiry. The first issue concerns the asymmetry in theory testing between psychology and physics and the resulting paradox that, because the psychological null hypothesis is always false, increases in precision in psychology always lead to weaker *tests of a theory, whereas the converse is true in physics. The second issue, related to the first, regards the slow progress observed in psychological research and the seeming unwillingness of social scientists to take seriously the Popperian requirements for intellectual honesty. We propose a good-enough principle to resolve Meehl's methodological paradox and appeal to a more powerful reconstruction of science developed by Lakatos (1978a, 1978b) to account for the actual practice of psychological researchers.*

From time to time every research discipline must reevaluate its method for generating and certifying knowledge. The actual practice of working scientists in a discipline must continually be subjected to severe criticism and be held accountable to standards of intellectual honesty, standards that are themselves revised in light of critical appraisal (Lakatos, 1978a). If, on a metatheoretical level, scientific methodology cannot be defended on rational grounds, then metatheory must be reconstructed so as to make science rationally justifiable. The history of science is replete with numerous such reconstructions, from the portrayal of science as being inductive and justificationist, to the more recent reconstructions favored by (naive and sophisticated) methodological falsificationists, such as Popper (1959), Lakatos (1978a), and Zahar (1973).

In the last two decades psychology, too, has been subjected to criticism for its research methodology. Of increasing concern is empirical psychology's use of inferential hypothesis-testing techniques and the way in which the information derived from these procedures is used to help us make decisions about the theories under test (e.g., Bakan, 1966; Lykken, 1968; Rozeboom, 1960).

In two penetrating essays, Meehl (1967, 1978) has cogently and effectively faulted the use of the traditional null-hypothesis significance test in psychological research. According to Meehl (1978, p. 817), "the almost universal reliance on merely refuting the null hypothesis as the standard method for corroborating substantive theories [in psychology] is a terrible mistake, is basically unsound, poor scientific strategy, and one of the worst things that ever happened in the history of psychology." He maintained that it leads to a methodological paradox when compared to theory testing in physics. In addition, Meehl (1978) pointed to the apparently slow progress in psychological research and the deleterious effect that null-hypothesis testing has had on the detection of progress in the accumulation of psychological knowledge. The cumulative effect of this criticism is to do nothing less than call into question the rational character of our empirical inquiries. As yet there has been no attempt to deal with the problems raised by Meehl by reconstructing the actual practice of psychologists into a logically defensible form. This is the purpose of the present article.

The two articles by Meehl seem to deal with two disparate issues—null-hypothesis testing and slow progress. Both issues, however, are linked in the methodological falsificationist reconstruction of science to the necessity for scientists to agree on what experimental outcomes are to be considered as disconfirming instances. We will argue that the methodological paradox can be ameliorated with the help of a "good-enough" principle, to be proposed here, so that hypothesis testing in psychology is not rationally disadvantaged when compared to physics. We will also account for the apparent slow progress in psychological research, and we will take issue with certain (though not all) claims made by Meehl (1978) in this regard. Both the methodological and the progress issues will be resolved by an appeal to the (sophisticated) methodological falsificationist reconstruction of science developed by Lakatos (1978a), an approach with which Meehl is familiar but one he did not apply to psychology in his articles.

Copyright 1985 by the American Psychological Association, Inc. 0003-066X/85/$00.75
Vol. 40, No. 1, 73–83

Meehl's Asymmetry Argument

Let us develop Meehl's argument. It is his contention that improved measurement precision has widely different effects in psychology and physics on the success of a theory in overcoming an "observational hurdle." Perfect precision in the behavioral sciences provides an easier hurdle for theories, whereas such accuracy in physics makes it much more difficult for a theory to survive. According to the Popperian reconstruction of science (Popper, 1959), scientific theories must be continually subjected to severe tests. But if the social sciences are immanently incapable of generating such tests, if they cannot expose their theories to the strongest possible threat of refutation, even with ever-increasing measurement precision, then their claim to scientific status might reasonably be questioned. Further, according to this view of research in the social sciences, there can be no question of scientific progress based on the rational consideration of experimental outcomes. Instead, progress is more a matter of psychological conversion (Kuhn, 1962).

Let us look more closely at the standard practice in psychology. On the basis of some theory T we derive the conclusion that a parameter δ will differ for two populations. In order to examine this conclusion, we can set up a point-null hypothesis, H_0: $\delta = 0$, and test this hypothesis against the predicted outcome, H_1: $\delta \neq 0$. However, it has also been recognized (Kaiser, 1960; Kimmel, 1957) that another question of interest is whether the difference is in a certain direction, and so we could instead test the directional null hypothesis, H_0^*: $\delta \leq 0$, against the directional alternative, H_1^*: $\delta > 0$. In such tests, we can make two types of errors. The Type I error would lead to rejecting H_0 or H_0^* when they are indeed true, whereas the Type II error involves not rejecting H_0 or H_0^* when they are false. The conventional methodology sets the Type I (or alpha) error rate at 5% and seeks to reduce the frequency of Type II errors. Such a reduction in the Type II error rate can be achieved by improving the logical structure of the experiment, reducing measurement errors, or increasing sample size.

Meehl pointed out that in the behavioral sciences, because of the large number of factors affecting variables, we would never expect two populations to have literally equal means. Hence, he concluded that the point-null hypothesis is always false. With infinite precision, we would always reject H_0. This is perhaps one reason to prefer the directional null hypothesis H_0^*.

But Meehl then conducted a thought experiment in which the direction predicted by T was assigned at random. In such an experiment, T provides no logical connection to the predicted direction and so is totally without merit. Because H_0 is always false, the two populations will always differ, but because the direction in H_0^* is assigned at random, with infinite precision we will reject H_0^* half of the time. Hence, Meehl concluded "that the effect of increased precision . . . is to yield a probability approaching ½ of corroborating our substantive theory by a significance test, *even if the theory is totally without merit*" (Meehl, 1967, p. 111, emphasis in original).

Meehl contrasted this state of affairs with that in physics, wherein the usual situation involves the prediction of a point value. That which corresponds to the point-null hypothesis is the value flowing as a consequence of a substantive theory T. An increase in statistical power in physics has the effect of stiffening the experimental hurdle by "*decreasing* the prior probability of a successful experimental outcome if the theory lacks verisimilitude, that is, precisely the reverse of the situation obtaining in the social sciences" (Meehl, 1967, p. 113). With infinite precision, and if the theory has no merit, the logical probability of it surviving such a test in physics is negligible; in the social sciences, this logical probability for H_0^* is one half.

Perhaps another way of describing the asymmetry in hypothesis testing between psychology and physics is to note that, in psychology, the point-null hypothesis is not what is derived from a substantive theory. Rather, it is a "straw-man" competitor whose rejection we interpret as increasing the plausibility of T. In physics, on the other hand, theories that entail point-null statistical hypotheses are the very ones physicists take seriously and hope to confirm. If 0 is a predicted outcome of interest, and $\tilde{0}$ is its logical complement, then the depiction of null and alternative statistical hypotheses in the two disciplines can be written as follows:

For psychology: H_0: $\tilde{0}$
H_1: 0

For physics: H_0: 0
H_1: $\tilde{0}$

Hence the methodological asymmetry between psychology and physics has two important aspects. The first is that physicists devolve substantive point values from theories for their point-null hypothesis test, whereas psychologists test for the "straw-man" com-

An earlier version of this article was read at the 1983 meeting of the American Educational Research Association. The authors are grateful to Robbie Case, Joel R. Levin, and Leonard Marascuilo for reading earlier drafts, and to Crescent L. Kringle for her help with the manuscript.

Requests for reprints should be sent to Ronald C. Serlin, Department of Educational Psychology, University of Wisconsin, Madison, Wisconsin 53706.

petitor, 0. The second is that increased precision in physics gravely threatens a theory with refutation, whereas such precision in psychology decreases such a threat.

It should be emphasized that Meehl's argument is not with statistical testing as such, but with what we infer about substantive theories given statistical information. The appraisal of a substantive theory T, in both physics and psychology, entails some constraint on the population value of the statistical parameter μ. The entailed constraint on μ is very strong in physics, often being a point value or function form. In psychology, however, the constraints on μ are said to be very weak, implying that it lies only in a half-uncurtailed interval whose prior probability is 50%. Statistical precision tells us how adequately we have established the actual value of μ and with what degree of confidence we can arrive at conclusions regarding T.

The important question, however, is this: Having arrived at an estimation of μ, even with perfect precision, how does this affect the plausibility of T? Meehl would argue that the typical null-hypothesis test in psychology is so weak as to be worthless when passed. This is so not because of any uncertainty regarding the posterior estimation of the tested parameter's value, but because the prior parameter constraint is so flabby. This is not the case in physics, where the prior constraint on μ is precise. In other words, in both physics and psychology, statistical techniques allow one to infer the posterior value of μ with great certainty. But the initial constraint on μ in psychology is so weak that the test cannot speak meaningfully to the plausibility of T. This is the methodological paradox posed by Meehl in his 1967 article. Other, related problems were described in a subsequent article (Meehl, 1978). To these we now turn.

Progress in Psychology

In his more recent article, Meehl (1978) made a number of valuable points, but we shall consider only a few of them here. Meehl rightly criticized the common practice of comparing substantive theories, or rival tests of the same theory, by examining the probability levels associated with statistical tests. In a typical review article, according to Meehl (1978), one often finds a "nose count" of favorable and unfavorable experimental outcomes, with particular interest paid to the level of significance at which the null hypotheses were rejected, as if this made any difference. This, according to Meehl, is preposterous, because in the Popperian scheme of things, a single refutation is more compelling than a host of corroborations. Thus, "the whole idea of simply counting noses is wrong, because a theory that has seven facts for it and three facts against it is *not* in good shape, and it would not be considered so in any developed science" (Meehl, 1978, p. 823).

This is not the only problem, in Meehl's view, for psychological methodology. An additional problem concerns the extent to which the *modus tollens*[1] can be effectively directed at the substantive theory under test. According to the Popperian reconstruction of science, an intellectually honest researcher must specify in advance the experimental state of affairs he or she will accept as falsifying the theory. But a theory T is never directly under test. Rather, it is the theory T plus a set of auxiliary theories, and the *ceteris paribus*[2] clause that are concomitantly and jointly put to the test. Hence negative empirical results could never decisively refute a theory because a researcher could always implicate instead one or more of the auxiliary theories, or some element subsumed under the *ceteris paribus* clause, as being responsible for the "refutation."

Meehl maintained that although this phenomenon is qualitatively the same for both the hard and soft sciences, the problem is quantitatively more severe for the soft sciences. One reason is that independent testing of auxiliary theories is harder to carry out. A second reason is that there is no (it is claimed) intimate connection, no sense of derivability, in the social sciences between auxiliary theories and the substantive theory T. Thus, according to Meehl (1978),

> Almost nothing we know or conjecture about the substantive theory helps us to any appreciable degree in firming up our reliance on the auxiliary (A). The situation in which A is merely conjoined to T in setting up our test of T makes it hard for us social scientists to fulfill a Popperian falsifiability requirement—to state before the fact what would count as a strong falsifier. (p. 819)

Later, however, Meehl seemed to suggest that the difficulty that psychology has in pointing the arrow of the *modus tollens* at the heart of a theory dooms not only rational psychological inquiry but the Popperian reconstruction of science as well. Meehl (1978) wrote that

> it is perhaps worth saying . . . that the above described situation . . . may represent a social fact about the way science works that presents grave difficulties for the Pop-

[1] The *modus tollens* is the logical form of Popper's (1959) falsification criterion, which permits the deductive testing of theories. The *modus tollens* can be represented in the following way: A implies B; not B; therefore, not A.

[2] In the conduct of an experiment it is assumed that contaminating or perturbing influences are not present and that the only effects to be observed are those intended by the experimenter. This assumption is conjoined to every hypothetical deduction from theory, and, in the technical language of the philosophy of science, is called the *ceteris paribus* clause. The *ceteris paribus* clause states that, for the sake of a particular experiment, all things are equal. We employ such terms for economy of exposition.

> perian reconstruction. That is, the stipulation beforehand that one will be pleased about substantive theory T when the numerical results come out as forecast, but will not necessarily abandon it when they do not, seems . . . about as blatant a violation of the Popperian commandment as you could commit. . . . But it seems in accordance with much scientific practice. (p. 821)

Hence, in his two articles, Meehl presented the behavioral scientist with two significant and related problems. With ever-increasing measurement precision we subject our substantive theories to ever-decreasing danger of refutation—we lower the "observational hurdle" that theories are required to surpass. In addition, accounting for progress in psychology is made difficult not only by the asymmetry just noted, but also by the practice of reviewers in valuing corroboration over refutation and by the difficulty of subjecting psychological theories to the threat of refutation via the *modus tollens.* We will attempt in the remainder of this article to resolve these difficulties and, perhaps, restore a sense of rationality to psychological inquiry.

The Lakatosian Reconstruction of Science

It was Lakatos's (1978a) contention that the comparability of scientific results and the assessment of scientific progress must take on a historical character, as we shall see. Lakatos phrased his own metatheory of science similarly by placing it in the historical context of prior reconstructions of science. In order to appreciate the progress made by Lakatos, it will prove useful to briefly recount part of this history as Lakatos described it. This is done not only to introduce the Lakatos model but also to prepare the groundwork for our defense of progress in psychological inquiry and our reconstruction of Meehl's argument concerning it.

We pick up the history with a consideration of several variants of methodological falsificationism. Falsificationist reconstructions attempt to account for the fact that scientific practice cannot be rationally defended if it is portrayed as proceeding inductively and that all scientific theories are hence equally unprovable (justificationism) and improbable (neojustificationism or probabilism). Methodological falsificationism holds that although science cannot prove theories or establish their probability via a probability calculus, it can, with the *modus tollens,* certainly disprove them. According to *dogmatic* falsificationism, once a theory is disproved and a refuting instance is uncovered, it should be eliminated from the corpus of scientific theories. Although all theories are said to be fallible, this reconstruction assumes that there exists an infallible empirical basis. That is, a demarcation is said to exist between facts and theories such that one could always unequivocally appeal to hard "facts" in the evaluation of fallible "theories." According to Lakatos, however, the dogmatic falsificationist position is untenable for two reasons. First, what are considered facts are only acceptable to us if we believe in certain theories that describe how our measuring instruments work. Second, as Kuhn (1962) and others (e.g., Polanyi, 1958) have pointed out, there is no strict psychological demarcation between fact and theory. Indeed, according to Lakatos (1978a) "there can be no sensations unimpregnated by expectations, and *therefore there is no natural demarcation between observational and theoretical propositions*" (p. 15, emphasis in original).

Yet, a more significant problem exists for the dogmatic falsificationist view—no theory can ever forbid any possible experimental outcome, in that negative results can always be attributed to other extraneous factors thought to be influential or to factors not previously taken into account (via the *ceteris paribus* clause). According to Lakatos (1978a),

> some scientific theories forbid an event occurring in some specific spatio-temporal region only on the condition that no other factor . . . has any influences on it. But then such theories never alone contradict a "basic" statement: they contradict at most a conjunction of a basic statement describing a spatio-temporally singular event and of a universal nonexistence statement saying that no other relevant cause is at work . . . [thus] the dogmatic falsificationist cannot possibly claim that such universal nonexistence statements belong to the empirical basis. (p. 17)

In other words, every scientific theory contains a *ceteris paribus* clause, so that it is a theory plus the *ceteris paribus* clause that is subjected to a test. Because it is always possible to replace the *ceteris paribus* clause, any single test of a theory is of little consequence. As Lakatos (1978a) pointed out, this seems to lead to the uncomfortable conclusion that all scientific theories are not only equally unprovable and improbable, they are all equally undisprovable as well!

Popper's (naive) methodological falsificationism attempted to rescue empirical science from skepticism by demonstrating that science is not only a corpus of assertions but also a system of conventions. The process of testing our theoretical conjectures is impossible without making a series of methodological decisions. Because there are no pure observations, what we regard as facts must be conventionally agreed upon in light of a " 'relevant technique' such that 'anyone who has learned it' will be able to *decide* that the statement is 'acceptable' " (Lakatos, 1978a, p. 22). Thus the potential falsifiers of a theory are granted "observational" status by decision. The truth value of such observations is arrived at by a relevant experimental technique. The (naive) methodological falsificationist appreciates the fact that experimental techniques and scientific theories (not necessarily mutually exclusive) are fallible, but, by decision, he or she assumes that such theories constitute unproblematic background knowledge to be subsumed by the *ceteris paribus* clause. Although this solves the problem of how to demarcate fact from theory, we are still left with the problem of how, given the *ceteris paribus* clause, we are able to subject a specific theory to refutation. Popper maintained that we do this by making yet another methodological decision. Researchers decide, before experiments are conducted, what state of affairs they will accept as a falsification of their theories, irrespective of *ceteris paribus.* This is a significant point not mentioned by Meehl. Popper was aware that theories could not be subjected cleanly to a test because of *ceteris paribus* and auxiliary theories. But he believed that to play the game fairly a researcher must specify what events would be considered falsifying without ad hoc appeals to *ceteris paribus.* Thus, the Popperian reconstruction is not endangered by the "auxiliary-theory" phenomenon in theory testing, as Meehl implied; it is simply handled by making a public, conventional decision to consider a theory falsified given the observation of a specified state of affairs.

Yet, in Lakatos's view, the Popperian reconstruction was still not sufficient, for it did not account for the tenacity with which theories are held in the face of seemingly disconfirming evidence. Lakatos's reconstruction, sophisticated methodological falsificationism, attempted to address this issue. In short, one must distinguish between criticism of a theory and its abandonment (see Lakatos, 1978a).

One may criticize a theory (or research program) by pointing out the existence of empirical anomalies, phenomena that the program claims to account for but does not. Mere criticism, however, mere refuting evidence, is never sufficient in itself to falsify a theory. This is so because researchers protect the "hard core" of the theory from refutation with a "protective belt" of auxiliary theories. This is the "negative heuristic" of a research program. The negative heuristic is the result of a methodological decision to cordon off the core of a theory from the threat of refutation. That is, we forbid the *modus tollens* to be directed at the hard core but insist instead that the auxiliary theories bear the brunt of the tests.

In addition to the negative heuristic, which tells us what path *not* to pursue, research programs also have a positive heuristic, which tells us the direction that research *should* take. It consists of models or suggestions on how to modify the "refutable" protective belt. The positive heuristic proceeds in the face of counterevidence and refutation, and there is no need to consider the presence of empirical anomalies as being decisive. These "refutations" are not ignored, but they are considered inconclusive until some later time when the positive heuristic must confront the disconfirming evidence and, it is hoped, turn it into supporting evidence. This reconstruction thus accounts for the relative autonomy and tenacity of theoretical science.

One should be prepared to abandon a research program, on the other hand, only if certain criteria are met. There must exist a rival research program that is powerful enough to account for all the facts of the former program. In addition, and importantly, the rival research program must anticipate new "facts," some of which have been corroborated. But even these criteria, although providing necessary grounds for inducing the abandonment of a research program, are not in themselves sufficient as long as the former program is "progressive," that is, as long as its positive heuristic is still capable of generating novel facts. Even if a research program is "degenerating," one is still entitled to embrace it so long as no rival program exists that satisfies the above criteria.

With this reconstruction of scientific methodology by Lakatos (1978a), we are now in a position to reexamine certain of Meehl's complaints against orthodox psychological research. Meehl was dissatisfied with reviewers who "counted noses," who valued corroborations instead of refutations, and who inevitably concluded that "further research is needed to explain the discrepancies" (Meehl, 1978, p. 822). Meehl considered this procedure lamentable and preposterous, but it is instead quite rational. It is not appropriate to speak of a theory being falsified. There can be no "refutation" until the emergence of a more powerful theory. Instead of the older views of science (dogmatic and naive falsificationism) that confront "theories" with "facts," with the only interesting outcome being progress through disconfirmation, the Lakatosian view pits *theories* against each other, with confirmations also providing outcomes of interest. Meehl was certainly correct when he deplored the practice of comparing theories by reference to probability levels, but he was not correct when he asserted that a theory that has "seven facts for it and three facts against it is *not* in good shape" (Meehl, 1978, p. 823). Indeed, the history of physics is testimony to the fact that subsequently successful research programs typically have proceeded in the face of "oceans of anomalies" and disconfirming evidence (Lakatos, 1978a). It is too rash to overthrow a theory because of a recalcitrant fact, and no one does so, either in psychology or in any of Meehl's "developed sciences." Instead, the rational procedure is to do what Meehl found proposterous, to examine empirical discrepancies by thoroughly testing the *ceteris paribus* clause. Even this is not always necessary when the program's positive heuristic is busy unearthing new facts, that is, when the research program is still "progressive." And even should a theory have seven facts against it and only three for it, one may still legitimately pose the theory, because, in the Lakatosian scheme of things, there can be no refutation prior to the emergence of a superseding theory satisfying the criteria noted above. Hence, the appraisal of theories and the detection of progress in science takes on a historical character, with rejection involving a multiple relation among competing theories.

An understanding of Lakatos's emphasis on research programs does allow us to be critical of experiments that are performed without reference to theoretical propositions. But Meehl (1978) suggested that such "naive guessing" constitutes the majority of social science research when he wrote,

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

The implication is that there are no research programs in psychological research. If this were true, then according to the above philosophy, such research would be truly unscientific. But the presence of research programs is not uncommon in the soft sciences. There have been a number of attempts recently to conceptualize psychological research along Lakatosian lines (Beilin, 1983; Lapsley & Serlin, 1984; Rowell, 1983; Urbach, 1974a, 1974b). Common to these case studies is the attempt to identify hard-core assumptions of the research programs in question and, in addition, the heuristic machinery that guides further development of the programs. Perhaps the most comprehensive case study has been provided by Urbach in his analysis of scientific progress and degeneration in the "IQ debate" between the hereditarian and environmentalist research programs. According to Urbach (1974a, p. 102), the hard core of the hereditarian research program consists of the twin propositions that (a) the cognitive activity of all individuals is related to "general intelligence" and (b) that individual and group differences in general mental capacity are the result of inherited differences. The heuristic of this program, according to Urbach, is embodied in at least two "methodological directives." The first directive is "to construct ever-improving tests of 'general intelligence' and to check these tests by using them to measure the IQs of people whose genetic relations are known from Mendelian theory" (Urbach, 1974a, p. 105). The second directive is to compare group differences in intelligence and also to investigate the relation between intelligence and other variables.

For the environmentalist position, the hard core consists of the proposition that the genetic inheritance for intelligence is constant for all individuals and that observed differences in intelligence are the result of environmental effects. Its heuristic contains the directive to seek those environmental factors said to mediate group differences in intellectual ability. In the course of developing these research programs in accordance with their respective methodological directives (i.e., heuristics), anomalies will surface, and disconfirming instances will be encountered. Appeals will be made to variables subsumed under *ceteris paribus,* variables previously relegated (by convention) to unproblematic background knowledge, in order to account for anomalies. Auxiliary "protective belt" hypotheses will be proposed and tested. In the hereditarian case, for example,

> If . . . tests do not yield the predicted pattern, then the hereditarian first blames the test. The test is declared "badly administered." For example, it may be conjectured that the subjects were not put sufficiently at ease during the test. If no such assumption succeeds in dissolving the anomaly, the test is declared "culture biased" against some people—in other words it is said to favor those people who possess some specialized knowledge or experience. In order to test this assumption, a new test must be employed which can be seen to exclude the putative, unfair cultural element. One can, of course, introduce more substantial changes in order to account for anomalies. (Urbach, 1974a, p. 106)

Clearly, when scientific activity is viewed in terms of research programs, there can be no "instant rationality" in the assessment of progress. Rather, assessment must take on a historical character. A research program is progressive so long as each new theory within the program not only accounts for the anomalies of its predecessor but anticipates novel facts as well, some of which have been corroborated. If successive theories within the program account *only* for past anomalies, then the program is degenerating (given a well-tested *ceteris paribus* clause). In the "IQ debate," for example, Urbach (1974a) contended that the environmental appeal to theories regarding socioeconomic, cultural, and personality factors constituted only ad hoc explanations of IQ differences and hence contributed to the degeneration of the environmentalist program.

Other case studies of Lakatosian research programs in psychological research can be found in the developmental literature. Lapsley and Serlin (in press), for example, have conceptualized the cognitive developmental approach to moral judgment in Lakatosian terms. The hard core of the cognitive developmental approach, according to Lapsley and Serlin, is the proposition that the structure of moral cognition must show stagelike development. The positive heuristic consists of the "suggestion" to deploy ever more powerful stage models until empirical realities are accounted for. This is certainly evidenced in the cognitive developmental literature by the proliferation of stage models that attempt to account for the complexity of structural development (e.g., Bickhard, 1978; Campbell & Richie, 1983; Flavell, 1972; Levine, 1979; Rest, 1979; see Puka, 1982, for a discussion of alternate stage models in moral judgment research). A historical consideration of this research program reveals not only theoretical revisions via protective belt hypotheses (e.g., Gibbs, 1979; Kohlberg, 1973; Murphy & Gilligan, 1980) but also the extension of stage models into other positive justice domains (e.g., Damon, 1975; Lapsley & Madar, 1983). On this latter basis, Lapsley and Serlin argued that the cognitive–developmental approach is a progressive and not a degenerating research program.

These case studies and others (Rowell, 1983) clearly indicate that psychological inquiries can be reconstructed as constituting various research programs where metatheoretical evaluative criteria can be applied with profit. When psychological research is so reconstructed, many of the problems raised by Meehl (1978) concerning the detection of progress are resolved. We have argued that certain practices condemned by Meehl, such as nose counting experimental outcomes and the valuation of corroboration, are not violations of intellectual standards of honesty under the Lakatosian model of science. Indeed, sophisticated methodological falsificationism emphasizes the cautious appraisal of competing theories, which necessarily involves repeated (and time-consuming) appeals to *ceteris paribus* and a historical consideration of experimental data, where corroborations also provide outcomes of interest. Meehl was perhaps too expectant of rapid progress because of his faith in the Popperian reconstruction of science, which is conceived as involving a rapidly developing and continuing series of "conjectures and refutations" (Popper, 1968), where a single negative instance can purge a theory from further consideration. As we have seen, however, this model does not accord with actual practice in even our most cherished and developed of sciences—physics. Scientific theories are tenacious. They are developed in the face of recalcitrant evidence by the force of suggestion of methodological directives—the heuristic machinery—and are fortified by networks of auxiliary theories. Hence, deliberate progress in psychology is to be expected, and the seeming reluctance to overthrow a theory faced with recalcitrant facts does not call into question the rational character of psychological research. However, it remains to be seen whether the methodological asymmetry between physics and psychology can be resolved, because the inability to find such a resolution would constitute an indictment of empirical psychology.

The Good-Enough Principle

How, then, shall we address the other important problem posed by Meehl, that concerning the poverty of traditional null-hypothesis testing in psychology? The answer, we feel, lies in adopting a methodology that is consistent with the previously described sophisticated falsificationism and that, even with infinite sample size, does not always reject the null hypothesis. It is a methodology already used by scientists, in both the hard and soft sciences, and for which statisticians (Hodges & Lehmann, 1954; Walster & Cleary, 1970) have already provided some guidelines. As Meehl (1978) has pointed out, when a scientist examines an experimental result, he or she "looks at the agreement, and comments that 'the results are in reasonably good accord with theory.' " That is, such a scientist has set standards that indicate what kinds of experimental outcomes are "good enough." This, in effect, imposes a set of constraints on the statistical parameters that we estimate from sample data. It is an extension of Popper's demand that scientists establish, in advance, what they will accept as a falsifying instance.

Let us see how this principle can be used to eliminate the problem in the soft sciences of infinite precision always rejecting null hypotheses. First consider the point-null case. A psychologist, in the test of an alternative hypothesis of interest, makes the null prediction that a treatment will have zero effect. Hence, the null hypothesis states that a particular variable, δ, possesses a certain expected value, which is 0. But because no theory is absolutely true, the value of δ can never be exactly equal to the theoretical value, 0. Therefore, a good-enough belt of width Δ must also be included in the prediction, so that a value $0 \pm \Delta$ is predicted. The value of Δ is chosen in advance and reflects the state of the art or the error in the best "known experimental technique" in the field. When the experiment is performed, a statistical test is applied to determine if δ is in the range $0 \pm \Delta$. If the data indicate that δ seems to be in the good-enough belt, then the complex null hypothesis cannot be rejected. But now the effects of increased sample size are not problematic, because with increased precision, the imprecision involved in estimating the population value is reduced. In the limit of infinite precision, one finds theoretical support simply by finding the sample value to be outside the range $0 \pm \Delta$. Thus, even with an infinite sample size, the point-null hypothesis, fortified with a good-enough belt, is not always false.

We hasten to add that precisely the same state of affairs obtains as well in physics. The hard scientist predicts from a substantive theory that a particular parameter will have a certain numerical value, say 2. But nature is just as unkind to physicists as it is to psychologists, in that the true value will never be exactly equal to the theoretical value. Hence, in a fashion similar to before, a good-enough belt is included in the prediction, yielding a value 2 ± Δ. An inference is made about the true parameter estimated by the data. If this parameter does not seem to be in the predicted range 2 ± Δ, one concludes that the empirical fit is not good enough, a methodology that parallels the psychological case. So although the asymmetry between psychology and physics is indeed real enough, it is only real in the sense that the nature of the point values tested under the null hypotheses are different. Without the good-enough belt, the null hypothesis will always be rejected if the sample size is infinite. Referring to the display comparing null hypotheses (see page 74), the rejection would be considered "support" for a psychological theory and a "disconfirmation" of a physical theory. In both cases, however, with an infinite sample size and without a good-enough belt, these results are known in advance. It is only with the aid of a good-enough belt that one can learn from a perfectly precise experiment in either discipline. Hence, both disciplines require the methodological decision to employ good-enough belts around parameters under test to avoid the paradoxical conclusions made inevitable by the prospect of infinite precision.

Let us now consider the more problematic case for theory testing in psychology, the directional null hypothesis. Under the aegis of the good-enough principle, one may not merely predict a direction. One also must specify in advance the magnitude of the change in that direction that is good enough. Thus, one would specify not only that a treatment will improve scores but also by how much the scores will increase. Again, if the statistical test indicates a possible increase less than that which is specified as good enough, the directional null hypothesis is retained. With infinite precision, one does not always reject the directional null hypothesis, and this is especially advantageous when the result is in the correct direction but only infinitesimally so.

It could still be countered that the good-enough methodology is ineffective when the directions are assigned at random. Consider the following possible results of a directional null-hypothesis test, where Δ is the good-enough quantitative prediction, and D the predicted direction (tail).

R_1: Δ true; D true

R_2: Δ true; D false

R_3: Δ false; D true

R_4: Δ false; D false

If the direction D were assigned at random, then the prior probability of rejecting H_0 under conditions R_1 or R_2 will be 50%, even with the employment of a good-enough principle. We would argue, however, that directional predictions are always an adjunct of quantitative, good-enough predictions and are never assigned at random. The prior probability of 50% holds only if the theory is totally without merit, a proviso that is crucial to Meehl's (1967) argument. This, of course, is never the case in either the behavioral or physical sciences. Our theories are not random conjectures bearing no logical relation to the directions deduced. Rather, one deduces the direction from theory. Meehl's (1967) example of irrationality in psychological research is compelling only because it describes a totally imaginary case.

But let us grant for the moment the force of Meehl's proviso. We would argue that even under conditions of random assignment of direction the good-enough principle provides outcomes of empirical interest. For example, it is important to note that with infinite precision and without a good-enough belt, the directional null hypothesis is rejected 50% of the time. This is so regardless of the true value of the population parameter. On the other hand, let us consider the directional null hypothesis with a good-enough belt, and let us examine the outcomes R_1 through R_4 under a condition of infinite precision. As noted earlier, under conditions R_1 or R_2, we reject H_0 50% of the time. Under conditions R_3 or R_4, however, we *never* reject the null hypothesis, even when the direction is assigned at random. Thus, the good-enough principle does stiffen the observational hurdle in the case of the directional null hypothesis and infinite precision.

In addition, the specification of a good-enough region generates an advance in theory building, according to Lakatosian principles. To illustrate, suppose gambling theorists know of a theory T_1 that correctly predicts the winners of athletic contests 50% of the time. Such a theory corresponds to one that satisfies Meehl's proviso that direction is randomly assigned. Let us further suppose that there is a competing theory, T_2, which correctly predicts not only the winners 50% of the time (direction) but also the point spread (good-enough belt). If one were to apply Lakatosian criteria to these theories, it is clear that T_2 is the better theory because it accounts for everything T_1 accounts for (direction, 50% of the time) and, in addition, accounts for more facts (point spread). Although both theories are deficient when it comes to direction, the one that employs a good-enough belt (T_2) is a credible advance of substantive interest to gambling theorists. Further, a theory, T_3, is almost certain to be developed that has the point-spread predicting power of T_2 and that does *not* assign directions at random. This theory will replace both T_1 and T_2 on Lakatosian principles.

There is another aspect of the good-enough principle that is of great importance. According to Neyman (1942), the statistical null hypothesis should be (by convention) associated with the empirical hypothesis for which false rejection is more serious than false acceptance. This is so because it is only the error of false rejection of the null hypothesis whose rate can be strictly controlled. The error of false acceptance has a rate that depends on the true parameter value. For a given false rejection error rate, the error of false acceptance can be minimized by increasing sample size, effect size, and precision. Because the error committed in falsely providing evidence in favor of a theory is considered a most grievous one, the null hypothesis should be associated with empirical evidence denying the truth of the theory. Thus, the common null hypothesis is made the complement of the theoretical deductions to be "proved."

But a more important reason for this complementarity is that the *modus tollens* can be directed only against the null hypothesis. Hence, logical considerations demand that, in order to conclude on the basis of evidence that an empirical fit to theory is good, one must set up a "straw-man" competitor to the theoretical deductions. If it is desired to conclude via the *modus tollens* that $\mu = \mu_0$, the appropriate null hypothesis must be $\mu \neq \mu_0$. But Bradley (1976) pointed out that once one sets up test criteria with a rate α of falsely supporting the theoretical point prediction $\mu = \mu_0$, the *maximum* probability of correctly supporting the theory is also α. This is untenable, and it is no doubt the reason that one rarely sees H_0: $\mu \neq \mu_0$ ever tested.

On the other hand, testing in this fashion with acceptable power is possible if one includes a good-enough belt in the null hypothesis. That is, the null hypothesis must state $\mu \leqslant \mu_0 - \Delta$ or $\mu \geqslant \mu_0 + \Delta$, in order to conclude upon rejection $\mu_0 - \Delta < \mu < \mu_0 + \Delta$. For such a null hypothesis, the probability of correct rejection can reach unity in precise experiments. Thus, it is only through the use of the good-enough belt that a statistical *modus tollens* can powerfully be aimed at a theory with a point prediction.

What, then, of asymmetry? We conclude that it is an oversimplification and that a physicist must set up the "straw-man" logical complement of the theoretical prediction, including a good-enough belt, in order to conclude that the fit of data to theory is good enough. But this is also what must be done in the soft sciences and for the same reasons. Hence, we see that the role of the good-enough principle is twofold: First, the specification of a good-enough region follows Popper's tenets in defining what the scientist will accept as "facts"; and second, it allows one to conduct a powerful test of a theory that makes a point prediction.

Recommendations

In the remainder of the essay we would like to suggest a statistical procedure to accompany good-enough hypothesis testing. The examination of effect size has often been suggested by statisticians as a concomitant measure to the significance value of an experimental outcome (Glass, McGaw, & Smith, 1981). Fisher's (1925) correlation ratio serves this purpose in analysis of variance. Hodges and Lehmann (1954) noted that a sufficient sample size can make small effects statistically significant and provided test criteria for various statistical procedures that allow one to test that a parameter exceeds a particular magnitude. Walster and Cleary (1970) offered methods for determining the appropriate sample size for detecting, with specified power, an important effect while guarding against detecting trivial effects. Thus, certain mechanisms are already in place for testing null hypotheses invoking a good-enough principle.

Let us examine a method that allows a test of a null hypothesis that includes a good-enough belt. We will illustrate this technique for the type of hypothesis encountered in the hard sciences, namely

$$H_0\colon |\mu - \mu_0| \geqslant \Delta,$$

where again Δ is our good-enough value. If we assume that the observations Y_i are independently normally distributed with true mean μ and variance σ^2, then the variable $\bar{Y} - \mu_0$ is normally distributed with variance σ^2/N, N being the sample size. Hence, $N(\bar{Y} - \mu_0)^2/S^2$ is distributed as a noncentral F with degrees of freedom 1 and $N - 1$ and noncentrality parameter $\lambda = N(\mu - \mu_0)^2/\sigma^2$. Under our null hypothesis, $\lambda \geqslant N\Delta^2/\sigma^2$. The error rate of false theoretical support, α, generated by incorrectly rejecting H_0, will become smaller as λ increases. Thus, following Mood and Graybill (1963, p. 296), we will construct the test criterion by setting $\lambda = N\Delta^2/\sigma^2$ and use the α percentile of the noncentral F distribution to set the critical value. The good-enough value must be specified in standard deviation units, a common practice in the behavioral sciences. As Bradley (1976) noted, the power of such a test ranges from α to 1, the lower limit attained when μ is barely within the good-enough region. If, however, we specify an "excellent fit" region, then we can calculate the sample size required to detect excellence with a prespecified power. If the excellent fit region is small, then the maximum power, calculated at $\mu = \mu_0$ (using the central F distribution) should be a reasonable approximation to the operating characteristics of the test.

For example, consider testing the hypothesis that the population mean is within 0.5 standard deviations of prediction, $\mu_0 = 10$. Then we have

$$H_0: |\mu - 10| \geqslant 0.5\sigma$$

versus

$$H_1: |\mu - 10| < 0.5\sigma.$$

Assume we had a sample of 32 observations, whose variance $S^2 = 1.2$ and whose mean $\bar{Y} = 9.8$. Then the test statistic would equal

$$F = \frac{32(9.8 - 10)^2}{1.2} = 1.07,$$

and we would reject H_0 if this F is less than the fifth percentile of the noncentral F distribution with degrees of freedom 1 and 31 and noncentrality parameter $\lambda = 32(.25) = 8$. This critical value equals 1.38, so that the sample value indicates that the population fit is "good enough." To approximate the power of such a test, we find that the probability of a central F variable, with 1 and 31 degrees of freedom, being less than the critical value is 0.75 (this is the maximum power).

The confidence interval associated with the good-enough principle illuminates an important feature of the method. The null hypothesis we have just tested is a statement concerning the noncentrality parameter. The confidence interval contains all values of the noncentrality parameter for which the null hypothesis would not be rejected. For the example above, we have, then, that

$$\frac{N(\bar{Y} - \mu_0)^2}{S^2} \geqslant F_{1,31,\lambda}(.05),$$

in order that H_0 not be rejected.This inequality holds for all λ satisfying

$$\lambda \leqslant 7.193,$$

so that

$$\frac{|\mu - \mu_0|}{\sigma} \leqslant 0.474$$

is the confidence interval. Such an interval emphasizes the fact that it is the closeness of prediction to the true mean that is of primary importance in establishing our fit to be good enough. Here the interval indicates not only that we are within the 0.5σ established by a hypothesis test but that we seem to be within 0.474σ of the true value.

It is also of interest to examine the asymptotic characteristics of this confidence interval. With infinite sample size, the confidence interval becomes $|\mu - \mu_0|/\sigma \leqslant |\bar{Y} - \mu_0|/S$. Hence, even with infinite sample size, the empirical fit to prediction can still be good enough. On the other hand, a central confidence interval reduces asymptotically to a single value, $\bar{Y}$, which we know in advance can never equal μ_0.

The purpose of this essay was to reexamine a number of methodological and procedural issues raised by Meehl (1967, 1978) that seemed to question the rationality of psychological inquiry. The first was in regard to the slow progress observed in psychological research and the seeming unwillingness of social scientists to take the Popperian requirements for intellectual honesty seriously. The second issue, related to the first, concerned the asymmetry in theory testing between psychology and physics and the resulting paradox that, because the psychological null hypothesis is always false, increases in precision (e.g., sample size) in psychology always lead to weaker tests of a theory, whereas the converse is true in physics. We have appealed to a more powerful (we think) reconstruction of science developed by Lakatos (1978a, 1978b) to account for the actual practice of psychological researchers, and we have proposed a good-enough principle to resolve Meehl's methodological paradox.

REFERENCES

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66,* 423–437.

Beilin, H. (1983). *Research programs and revolutions in developmental psychology.* Paper presented at the biennial meeting of the Society for Research in Child Development, Detroit.

Bickhard, M. (1978). The nature of developmental stages. *Human Development, 21,* 217–233.

Bradley, J. (1976). *Probability; decision; statistics.* Englewood Cliffs, NJ: Prentice Hall.

Campbell, R., & Richie, D. M. (1983). Problems in the theory of developmental sequences. *Human Development, 26,* 156–172.

Damon, W. (1975). Early conceptions of positive justice as related to the development of logical operations. *Child Development, 46,* 301–312.

Fisher, R. (1925). *Statistical methods for research workers.* Edinburgh, Scotland: Oliver & Boyd.

Flavell, J. (1972). An analysis of cognitive-developmental sequences. *Genetic Psychology Monographs, 86,* 279–350.

Gibbs, J. (1979). Kohlberg's moral stage theory: A Piagetian revision. *Human Development, 22,* 89–112.

Glass, G., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research.* Beverly Hills, CA: Sage

Hodges, J., & Lehmann, E. (1954). Testing the approximate validity of statistical hypotheses. *Journal of the Royal Statistical Society* (B), *16,* 261–268.

Kaiser, H. (1960). Directional statistical decisions. *Psychological Review, 67,* 160–167.

Kimmel, H. (1957). Three criteria for the use of one-tailed tests. *Psychological Bulletin, 54,* 351–353.

Kohlberg, L. (1973). Continuities in childhood and adult moral development revisited. In P. Baltes & K. Schaie (Eds.), *Lifespan developmental psychology.* New York: Academic Press.

Kuhn, T. (1962). *The structure of scientific revolutions.* Chicago: University of Chicago Press.

Lakatos, I. (1978a). Falsification and the methodology of scientific research programs. In J. Worrall & G. Currie (Eds.), *The methodology of scientific research programs: Imre Lakatos philosophical papers* (Vol. 1). Cambridge, England: Cambridge University Press.

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

Lapsley, D., & Madar, M. (1983). *Retributive justice reasoning in children.* Paper presented at the biennial meeting of the Society for Research in Child Development, Detroit.

Lapsley, D., & Serlin, R. (1984). On the alleged degeneration of the Kohlbergian research program. *Educational Theory, 34,* 157–169.

Levine, C. (1979). Stage acquisition and stage use: An appraisal of stage displacement explanations of variation in moral reasoning. *Human Development, 22,* 145–164.

Lykken, D. (1968). Statistical significance in psychological research. *Psychological Bulletin, 70,* 151–159.

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.

Mood, A., & Graybill, F. (1963). *Introduction to the theory of statistics* (2nd ed.). New York: McGraw-Hill.

Murphy, J., & Gilligan, C. (1980). Moral development in late adolescence and adulthood: A critique and reconstruction of Kohlberg's theory. *Human Development, 23,* 77–104.

Neyman, J. (1942). Basic ideas and theory of testing statistical hypothesis. *Journal of the Royal Statistical Society, 105,* 292–327.

Polanyi, M. (1958). *Personal knowledge.* Chicago: University of Chicago Press.

Popper, K. (1959). *The logic of scientific discovery.* New York: Basic Books.

Popper, K. (1968). *Conjectures and refutations.* London: Routledge & Kegan Paul.

Puka, B. (1982). Interdisciplinary treatment of Kohlberg. *Ethics, 92,* 471–492.

Rest, J. (1979). *Development in judging moral issues.* Minneapolis: University of Minnesota Press.

Rowell, J. (1983). Equilibration: Development of the hard core of the Piagetian research program. *Human Development, 26,* 61–71.

Rozeboom, W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin, 67,* 416–428.

Urbach, P. (1974a). Progress and degeneration in the "IQ debate" (I). *British Journal for the Philosophy of Science, 25,* 99–135.

Urbach, P. (1974b). Progress and degeneration in the "IQ debate" (II). *British Journal for the Philosophy of Science, 25,* 235–259.

Walstar, G., & Cleary, T. (1970). Statistical significance as a decision rule. In E. Borgatta & G. Bohrnstedt (Eds.), *Sociological methodology.* San Francisco: Jossey Bass.

Zahar, E. (1973). Why did Einstein's programme supercede Lorentz's? (I). *British Journal for the Philosophy of Science, 24,* 95–123.