6 Rational Appraisal of Psychological Research and the Good-Enough Principle

Ronald C. Serlin
University of Wisconsin-Madison

Daniel K. Lapsley
University of Notre Dame

INTRODUCTION

The extent to which psychological research can be suitably described as a scientific enterprise has long been a source of doubt, anxiety, and reflection. To those who are rightly impressed by the achievements in the physical sciences, the activity of psychological researchers seems inadequate and impoverished by comparison. In order to be described as a legitimate scientific enterprise it would seem that psychology must be able to satisfy some minimum standards of scientific adequacy. One should reasonably expect, for example, that psychology be capable of generating powerful theories that can be severely tested by precise hypothesis-testing procedures. It should further be expected that one be able to detect cumulative progress in the various theoretical domains of the discipline as a result of these theory-appraising methodologies. But the ability of psychological research to satisfy these twin expectations is precisely what is called into question. Indeed, psychological research is often said to be atheoretical, noncumulative, and saddled with a statistical hypothesis-testing methodology that is either deeply flawed or else ill-used by researchers.

Although there are numerous critics of significance testing (e.g., Morrison & Henkle, 1970; Spielman, 1974), Meehl (1967, 1978, 1986) effectively critiqued standard practices in the "softer" areas of social science research. He wrote, for example, that "the almost universal reliance on merely refuting the null hypothesis is a terrible mistake, is basically unsound, poor scientific strategy, and one of the worst things that ever happened in the history of psychology" (Meehl, 1978, p. 817). Meehl (1978) also faulted psychological research for its lack of cumulative progress. According to Meehl (1978), "It is simply a sad fact that in soft

psychology theories come and go more as a function of baffled boredom than anything else; and the enterprise shows a disturbing absence of that cumulative character that is so impressive in disciplines like astronomy, molecular biology, and genetics" (p. 807). The lack of cumulative progress, and the poverty of null hypothesis testing, are not unrelated problems, as will become evident. If they prove to be intractable problems for psychology then the very rationality of the research enterprise is legitimately called into question. That is, if these criticisms are sustained, if it is true that psychological theories cannot in fact be subjected to severe testing and do not cumulate into well-corroborated empirical knowledge, then the scientific character of research in psychology might well be a delusion. The purpose of this chapter is to reexamine these issues. After first examining the Meehlian complaints against psychological research in more detail, we then propose a number of remedies. Against the claim that the significance test cannot be made to threaten a theory with refutation, we propose a "good-enough" methodology that claims to do precisely that. Against the claim that psychological research is not cumulative, we argue, following Lakatos (1978a), that progress in research is never plainly evident but must instead be excavated from historical reconstructions of the various literatures. Along the way we provide examples of how one uses "good-enough" hypothesis testing. We also argue that the comparison with physics is not always to our disadvantage when the good-enough methodology and certain Lakatosian considerations are kept in mind. Finally, we conclude with a discussion of what rational appraisal of psychological research might look like, and how this might have an impact on graduate training in psychology.

THE MEEHLIAN INDICTMENT OF PSYCHOLOGY

The Methodological Paradox

In Meehl's (1967) view, improved measurement precision in the behavioral sciences has the paradoxical effect of yielding weaker tests of substantive theories. This paradox hinges on the fact that the psychological point null hypotheses are always false. Insofar as psychological variables are invariably contaminated by a large number of "crud" factors, one would never expect any two populations to have literally equal means. It follows, then, that one would always expect to reject the point null hypothesis if statistical power is sufficiently great. Hence, because the null hypothesis is always false, improved precision in the behavioral sciences provides an easier hurdle for theories to overcome, which violates the Popperian (Popper, 1959) tenet that theories must be subjected to severe tests.

The implication of this argument is quite startling. In order for a theory to be corroborated, under a standard account, it must face the heat of refutation. It

must be subjected to severe tests and be fairly confronted with the possibility of falsification. If a theory cannot be falsified, it fails the Popperian demarcation criterion as to what is to count as a meaningful scientific theory. Yet it would appear that psychological theories are not fairly confronted with the possibility of refutation. If the null hypothesis is always false, and if its rejection is thought to be an indication of the plausibility of the substantive alternative hypothesis, then it follows that one need never doubt the plausibility of the alternative hypothesis, because the rejection of the null hypothesis is guaranteed given sufficient power. But this hardly constitutes a severe test for the substantive theory, which suggests that the ritualistic rejection of null hypotheses and the corroboration of theories are two different matters in psychological research.

This state of affairs is contrasted with what is taken to be the standard case in physics (Meehl, 1967). According to Meehl (1967), the typical theory in physics predicts a point value or function form. That which corresponds to the point null hypothesis in psychology is a value derived as a consequence of a substantive theory. 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). If a physical theory has no merit, it will not survive an experimental test, given perfect precision. If a psychological theory has no merit, the logical probability of it surviving such a test is said to approach one.

The asymmetry in hypothesis testing between psychology and physics can be traced to the fact that in psychology, the point null hypothesis is not derived from a substantive theory. It is a "strawman" competitor whose rejection we interpret as increasing the plausibility of the substantive theory. But this interpretation is hazardous if the distinction between statistical hypotheses and substantive theories is to be respected (Bolles, 1962; Meehl, 1978). A substantive theory is a conjecture about the nature of psychological processes, entities, and phenomena. A statistical hypothesis is a conjecture about the value of a population parameter. If a null statistical hypothesis is not derived from substantive theoretical considerations, which seems the case in psychological research, then its rejection would not (necessarily) increase the plausibility of the substantive theory. Hence the chasm between statistical hypothesis and substantive theory is very wide in psychology. But in physics, theories that entail point predictions are the very ones that physicists take seriously and hope to confirm. The chasm between theory and statistical hypothesis is not terribly wide. Consequently, the asymmetry between psychology and physics has two features. First, physicists devolve subtantive point values for their statistical tests, whereas psychologists test for the strawman competitor, zero. Second, increased precision in physics gravely threatens a theory with refutation, whereas such precision in psychology decreases such a threat. It would appear, then, that significance testing in psychology is a vacuous exercise, neither corroborating nor refuting our substantive

theories. If Meehl is correct, then there would indeed be little reason for anyone to embrace or discard a theory for reasons other than curiosity, stubbornness or "baffled boredom." Consequently, it should not be surprising that psychological research seems to lack that cumulative quality that we tend to associate with more developed sciences. We now turn briefly to this topic.

Slow Progress in Psychology

As already noted, much of the difficulty concerning progress in psychological research is laid at the door of significance testing, which Meehl (1978, p. 806) described as "a poor way of doing science." Yet there are perhaps special reasons why cumulative progress is a more elusive achievement for psychological researchers. Meehl (1978), for example, catalogued 20 features of our subject matter that would seem to make it difficult to pin down psychological knowledge with any confidence. Of particular interest is the claim that psychological constructs are typically awash in a sea of context-dependent "stochastologicals," which refers to the fact that rather than dealing with lawlike relationships for which a certain amount of nomic necessity exists, we tend to deal mostly with "correlations, tendencies, statistical clusterings, increments in probabilities, and altered stochastic dispositions" (Meehl, 1978, p. 813), all of which show strong context dependence. Furthermore, Meehl noted that we rarely can know the complete list of relevant contextual influences. If we do know some of them, we can rarely specify the function form of the context dependency, nor the numerical values of the parameters for the function forms that we do know. This context dependence tends to make theory appraisal problematic. "When the observational corroborators of the theory consist wholly of percentages, crude curve fits, correlations, significance tests, and distribution overlaps, it is difficult or impossible to see clearly when a given batch of empirical data refutes a theory or even when two batches of data are 'inconsistent'" (Meehl, 1978, p. 814).

A similar indictment is leveled against the nature of psychological theories. Theories are said to be only loosely situated within a nomological network, such that auxiliary theories are just as problematic as the theory under test. In the Popperian view, a theory is never directly put to the test. Rather, it is the theory, plus a set of auxiliary theories that are jointly tested. Hence negative empirical results could never decisively refute a theory because a researcher could always implicate one of the auxiliaries as being responsible for the putative refutation. Although this describes the case with theory appraisal in the hard sciences, the problem is said to be more severe for the soft sciences because of the looseness with which theories are connected with auxiliaries within the network. Not only is independent testing of auxiliaries harder to carry out in psychology, it is also claimed that there is no intimate connection, no sense of derivability, between auxiliary and substantive theories in the first place. Thus, according to Meehl (1978):

> Almost nothing we know or conjecture about the substantive theory [T] helps us to an 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)

Hence it becomes clear, at least for Meehl, that slow progress is related to the fact that we cannot refute our theories, we cannot subject them to severe test via the *modus tollens*. And this is the result not only of certain peculiarities in significance testing (the asymmetry problem), but also because of the flabby nature of our theories, and the fact that we can never decisively bring data to bear on any one theory in particular.

We are now in a position to reconsider the Meehlian complaints against the standard practice of psychological researchers. We argue that a consideration of the Lakatosian reconstruction of science (Lakatos, 1978a) will take us some distance in restoring some semblance of rationality in our research practices, and in indicating how it might be possible to detect cumulative progress. We next describe a "good-enough" hypothesis-testing strategy that addresses the asymmetry problem noted by Meehl and allows a fortified test of a null hypothesis that is not always false.

SLOW PROGRESS RECONSIDERED: AN HISTORICIST APPROACH

One of the dogmas of positivism purported that there was a natural demarcation between observation ("facts") and theory. This view has long been discredited, as pointed out here. Although it is odd to hear anyone still try to maintain the distinction between facts and theories, it would be a mistake to assume that all vestiges of this dogma have been eliminated from our thinking about the scientific enterprise. Hardly anyone, for example, save the unreconstructed Baconian, would stare open-faced at nature waiting for the "facts" to make themselves evident. Rather, in order to make headway, in order to do science, one puts forward conjectures, one constructs theories, and within the context of theories, facts begin to emerge and to become *sensible*. We take it not to be controversial to assert that facts cannot exist outside the texture of theories. Yet if this formulation is readily accepted, as we take it to be, why should there also be an expectation that cumulative progress in scientific research is something that is plainly evident in the absence of suitable theory as to how science works? Just as one does not idly stare at nature waiting for "observations" to surface, one does not survey literatures waiting for progress to emerge unaided by theoretical considerations as to what "progress" amounts to. If evidence of progress is the "fact" that one is searching for, then one requires a theory that permits the excavation of the relevant data. One requires a theory that spec-

ifies how the history of science is to be reconstructed, by what criteria progress is to be ascertained, and the like. There are obviously numerous theories as to how science works, which suggests that, like all theories, conjectures about the scientific enterprise itself are fallible. But this need not lead to any thoroughgoing skepticism. Our approach has been to appeal to the work of Lakatos.

The Lakatosian Framework

According to Lakatos (1978a), the comparability of scientific results and the assessment of scientific progress must take on a historical character. When one takes a historical view of a research program one is often struck by the fact that theories are rarely abandoned. Indeed, theories are often tenaciously held even in the face of seemingly disconfirming evidence. Rarely, if ever, do scientists treat anomalies as falsifications of a theory. No theory is ever abandoned because of a "refuting" instance, because any negative result can always be attributed instead to extraneous factors that were provisionally treated as "unproblematic background knowledge," or to factors subsumed by the *ceteris paribus* clause that is (at least implicitly) appended to every theoretical deduction. Lakatos' (1978a) own view, which he called *sophisticated falsificationism,* attempted to account for this evident feature of science. In this view, one must distinguish between criticism of a theory and its abandonment. Mere criticism of a theory is never sufficient grounds for falsification. This is so because, in Lakatos' view, the "hard core" of a theory is protected from refutation by a "protective belt" of auxiliary theories, such that the arrow of the *modus tollens* must always be directed away from the core to the auxiliaries. Lakatos called this decision the *negative heuristic* of a research program. The negative heuristic specifies the path that the research program is not to take. That is, one decides to cordon off the core of the theory from the threat of refutation, insisting instead that the auxiliaries bear the brunt of the tests.

The direction that the research program should take is subsumed by the *positive heuristic* of the program. It specifies how one is to proceed in order to generate novel facts and thereby increase the empirical content of the program. It consists of models or suggestions on how to modify the refutable protective belt, on how to digest anomalies. Indeed, the positive heuristic of a research program bids one to proceed in the face of counterevidence and *refutation.* Empirical anomalies are never considered decisive, because "all theories are born refuted and die refuted" (Lakatos, 1978d, p. 5). Rather, these refutations are merely considered "inconclusive" until some later time when (it is hoped) the positive heuristic can turn the disconfirming evidence into corroborating evidence. In other words, the positive heuristic specifies how successive modifications and adjustments are to be developed within a research program so that anomaly and refutation are digested. Each successive modification yields a new theory, so the nature of scientific appraisal is shifted from the appraisal of isolated theories ("in

light of the evidence") to the appraisal of a *series* of theories (research programs), "where each subsequent theory results from adding auxiliary clauses to the previous theory in order to accommodate some anomaly" (Lakatos, 1978a, p. 33). A research program (or "series of theories") is said to be *theoretically* progressive if it has excess empirical content over its predecessor. It is said to be *empirically* progressive if some of the excess content is also corroborated. Furthermore, a research program is said to be *scientific* if it is at least theoretically progressive, and *pseudoscientific* if it is not. There will, of course, be anomalies encountered at every step in the development of a program. But we may rationally decide not to allow these putative refutations to transmit falsity to the hard core if the corroborated content of the protecting belt of auxiliaries increases, that is, if the research program, under the aegis of the positive heuristic, is theoretically and empirically progressive. This account of the scientific enterprise clearly explains the relative autonomy of theoretical science and the tenacity by which theories are embraced even in the face of *counterevidence*. The methodology of research programs, according to Lakatos (1978c),

> is more tolerant [than naive falsificationism] in the sense that it allows a research program to outgrow infantile diseases, such as inconsistent foundations and occasional ad hoc moves. Anomalies, inconsistencies, and ad hoc strategems can be consistent with progress. [But the appraisal of research programs] is also more strict in that it demands not only that a research program should successfully predict novel facts but also that the protective belt of its auxiliaries should be largely built according to a preconceived unifying idea, laid down in advance in the positive heuristic of the research program. (p. 149)

But when should one abandon a theory? On the Lakatosian account, only when certain criteria are met. There must exist a rival program that is powerful enough to account for all of the facts of the former program and, importantly, also possess sufficient generative power to anticipate novel facts, some of which have been corroborated (Lakatos, 1978a). Although one is entitled to embrace a rival under these conditions, they do not constitute sufficient grounds as long as the former program is still *progressive,* that is, as long as its positive heuristic is still capable of anticipating novel facts. Lakatos (1978a) also noted that one may still cling to a *degenerating* research program so long as no rival program exists that satisfies the aforementioned criteria.

There are a number of implications of this view that should be noted. First, the empirical character of a scientific theory and scientific growth are mutually defining. What gives science its scientific character is not the fact that it generates theories capable of making testable assertions about the nature of reality, but rather the fact that successive theories generated by a science possess excess content and excess corroboration when compared against predecessor or competing theories (Lakatos, 1978b; Serlin & Lapsley, in press). It is the *progressive*

character of a research program and the *growth of knowledge* that it represents, that determines whether or not a program is scientific. Indeed, a theory is never eliminated because it fails a test. The fate of a theory never depends on the results of experiments. Rather, *falsification* depends on the emergence of better theories, where "better theories" anticipate novel facts (excess content), some of which have been corroborated (excess corroboration). If a theory possesses these features, that is, if it represents growth, it is deemed *scientific*. If a theory does not contribute to growth, if it is ad hoc in one of the several senses described by Lakatos (1978a, see footnote 1, p. 88), it is deemed *pseudoscientific*. Consequently, in Lakatos' view, there can be no "instant rationality" in the appraisal of a research program, but rather appraisal must be guided by a consideration of the historical record of empirical successes and failures that are seen in the light of the track record of rival theories. That is to say, "there are no such things as crucial experiments, at least not if these are meant to be experiments which can instantly overthrow a research program. In fact, when one research program suffers defeat and is superceded by another, we may, with long hindsight, call an experiment crucial if it turns out to have provided a spectacular corroborating instance for the victorious program and a failure for the defeated one" (Lakatos, 1978a, p. 86).

With this brief review of the Lakatosian perspective we are now in a position to reexamine the slow progress issue in psychological research. In our view, as noted earlier, the progressive and cumulative character of a research program is not a self-evident fact that can be known in the absence of a theoretical perspective that directs our analysis of a literature. The "methodology of scientific research programs" articulated by Lakatos provides one way to conduct this analysis. It insists that evidential support be considered a comparative, historical matter, one that is to be excavated and *reconstructed* from the history of science. If the force of this analysis is granted, then one sees that slow progress is one of the defining features of science, and not an indictment. Indeed, "sophisticated falsificationism is a slower but possibly safer process" (Lakatos, 1978a, p. 40) than other theories of science (e.g., naive falsificationism), where there is an expectation that the content knowledge of science should grow linearly by means of a rapid succession of "conjectures and refutations" (Popper, 1963). But as we have seen, refutations are never decisive. When guided by a positive heuristic a theory forges ahead in almost complete disregard for refutations, for theories are tenaciously held even in the face of disconfirming evidence.

Consequently, if one of the complaints against psychological research is that we cannot seem to refute our theories, then perhaps the adoption of a Lakatosian perspective helps us see why this is an unreasonable expectation in the first place. It takes a long time to appraise a research program. Appraisal is not instant, based on the results of isolated experiments and the discovery of anomaly. Rather, as Lakatos (1978a, p. 35) pointed out, "There is no falsification before the emergence of a better theory. . . . Falsification is not simply a relation

between a theory and the empirical basis, but a multiple relation between competing theories, the original 'empirical basis,' and the empirical growth resulting from the competition. Falsification can thus be said to have a '*historical character*'" (p. 35).

Unfortunately, the appraisal of psychological theories and the search for cumulative growth is rarely conducted under the aegis of the "methodology of scientific research programs." Meehl (1967) was correct in his criticism of the habit of literature reviewers, who seem content merely to "count noses," that is, merely to tally the empirical successes and failures of a research program. This is a defective way to review a literature not because (according to Meehl) reviewers undervalue refutations, but rather because it is done in a way that is *ahistorical.* The typical literature review, particularly of the narrative sort, is too expectant of "instant rationality," and therefore neglects the historical comparison against rivals and the requirements for continuous growth. As Feldman (1971, p. 86; see also Rosnow & Rosenthal, 1989) noted, defecient literature reviewing "might account in part for the relatively unimpressive degree of cumulative knowledge in many fields in the behavioral sciences." When reviewing is done from the perspective of the Popperian falsificationist, who expects theories to be overthrown given the slightest hint of recalcitrant evidence, then it is little wonder that "slow progress" and "lack of cumulative findings" should become an indictment of psychological research.

Fortunately, a number of papers have recently appeared that would seem to address the problem of literature reviewing in the social sciences (e.g., Cooper, 1989; Jackson, 1980; Ladas, 1980), and the emergence of meta-analysis as a tool for evaluating scientific literatures is of considerable significance (e.g., Hedges & Olkin, 1985). There are also a number of instances of the use of the methodology of scientific research programs in the literature that provide ready examples of how the methodology might be used to detect cumulative progress. Case (1985), for example, provided a masterful reconstruction of the cognitive development research program. Lapsley and Serlin (1984) and Phillips and Nicolayev (1978) subjected the Kohlbergian research program to such an analysis, and Urbach (1974a, 1974b) charted the progress and degeneration in the IQ debate.

Nothing that has been said thus far should be construed as absolving research psychologists of all of the indictments that are typically leveled against them. Meehl (1967) was undoubtedly correct when he complained about the ad hoc nature of much of what passes as theory-building in some areas of the social sciences. Lakatos (1978a) distinguished between three kinds of ad hoc propositions. What he called *ad hoc*$_1$ are those propositions that do not increase the novel content of a theory. Theories are *ad hoc*$_2$ when they propose novel content, but none of which is corroborated. *Ad hoc*$_3$ strategems achieve "progress" with a "patched up, arbitrary series of disconnected theories" (Lakatos, 1978a, p. 88). Although all three kinds of ad hoc strategems are to be found in even the most mature sciences, that ad hoc$_3$ maneuvers are overrepresented in at least some

domains of psychological research seems hard to deny. We would agree with Lakatos (1978a, see footnote 4, p. 89) that the "methodology of research programs," if rigorously adopted by researchers and reviewers, would go a long way to stem this form of "intellectual pollution."

Although the foregoing discussion explains the apparent lack of progress in psychology, we have not yet dealt with Meehl's asymmetry problem. If this issue is not successfully addressed, then there is no way for psychologists even to test their theories rationally, let alone appraise them via a Lakatosian perspective. That is, the problem is not so much that we inappropriately appraise the successes and failures and therefore fail to detect progress, but rather that our tests are uninformative either way, given the deficiencies noted by Meehl with regard to significance testing. To resolve these difficulties, we introduce a *good-enough principle*, which allows one to obtain usable information by fortifying the null hypothesis. This enables significance tests to provide stiff observational hurdles for theories to overcome.

THE GOOD-ENOUGH PRINCIPLE

The asymmetry problem posed by Meehl is solved by examining Meehl's (1978) own account of actual scientific practice. Based in part on this account, we propose a good-enough principle and describe its associated statistical methodology. The suggested methodology does not result, even with an infinite sample size, in an inevitable rejection of the null hypothesis, and it is based on Popper's demand that scientists agree, in advance, to what they define as a falsifying experimental outcome. Although we have briefly described the methodology elsewhere (Serlin & Lapsley, 1985), we feel that a detailed account here will prove useful.

As Meehl (1978, p. 825) pointed out, when a scientist evaluates the results of an experiment he or she "looks at the agreement, and comments that 'the results are in reasonably good accord with theory.'" Such a scientist has set Popperian standards that indicate what kinds of experimental results are *good enough*, so that the experiment would not be considered a falsification. By implication, the standards would also reveal when a falsification had occurred.

It is important to point out that it is because all theories are false that a perfectly precise experimental outcome will *never* exactly agree with theoretical prediction. In order for psychologists to conduct and evaluate experiments whose outcomes are not known in advance, a good-enough belt around the theoretical prediction must be employed. This requirement is necessary even for theoretical predictions that are point values or function forms.

The Popperian implications of the good-enough principle for the "straw man" point null hypothesis are direct. Let us say that a psychological theory predicts that the means of two populations will differ on some measure. The typical point

null hypothesis would then state that the two population means are equal. But like all other theories, the conjecture posited by the point null hypothesis is false. This results in two problems. First, because the point null hypothesis is false, a large enough sample size will always lead to its rejection. And second, because the logical complement to the theoretical prediction comprises the null hypothesis, this inevitable rejection will lead to theoretical "support."

These deficiencies, however, can be corrected. First, a good-enough belt must surround the point value of zero, so that the theoretical prediction would now correctly state that the two population means differ by more than some good-enough value, say Δ_S (here, Δ_S represents the *smallest* difference that would constitute a nontrivial effect). And second, the appropriate null hypothesis that is *derived* from this theoretical prediction would state that the two population means differ by Δ_S or less. In this way, with perfect precision, a null hypothesis that tests a theory that is good enough would always be rejected, and a null hypothesis testing a theory that is not good enough would never be rejected. The logic of the *modus tollens* at work here now parallels that used in physics, in that the theoretical prediction logically yields the conclusion to be drawn, and a rejection of the appropriate null hypothesis allows one to conclude that the substantive theory has been confirmed.

Thus, the good-enough principle allows psychologists to specify in advance of an experiment what they would consider to be a falsifying instance. A Popperian corollary to the good-enough principle states that the researcher must also specify in advance the Type I error rate to be allotted to the hypothesis test. As with all Popperian specifications, the Type I error rate assignment is a methodological decision that must be endorsed in order to put a theoretical proposition to the test. Allowing the Type I error rate to "float" would mean that even after the experiment a result would be uninformative as to the success or failure of the theoretical prediction. In this context, Cohen (1990, p. 1311) voiced an "amen" to Rosnow and Rosenthal's (1989, p. 1277), "Surely, God loves the .06 nearly as much as the .05." We merely wonder what She thinks about the .07?[1] At some point the researcher must decide on an appropriate Type I error rate, and it is Popperian to make the decision in advance of the experiment.

In addition, similar considerations apply to testing psychological point value or function form predictions. Let us assume here that a psychological theory predicts that two population means do not differ, or they differ by a specific amount, or the relationship in the population between two variables is linear. Using only the first example (the others follow parallel logic), the prediction that the means differ by zero will never be supported by a perfectly precise experiment, so that the prediction must be modified to state that the means will differ by less than some good-enough value, say Δ_M (here, Δ_M represents the *most* that the means should differ). The null hypothesis derived from this theoretical pre-

[1]This sentiment was originally uttered by Mike Seaman, University of South Carolina.

diction would state that the population means differ by Δ_M or more. With the Type I error rate decided in advance, a rejection of this null hypothesis would constitute a confirming instance for the theory.

It may seem that this emphasis on confirmation is contrary to the Popperian emphasis on falsification. This is not the case. First, as Lakatos (1978a) noted,

> Our considerations show that the positive heuristic forges ahead with almost complete disregard of "refutations": it may seem that it is the "*verifications*" rather than the refutations which provide the contact points with reality. Although one must point out that any "verification" of the n + 1-th version of the program is a refutation of the n-th version, we cannot deny that *some* defeats of the subsequent versions are always foreseen: it is the "verifications" which keep the program going, recalcitrant instances notwithstanding. (pp. 51–52)

And second, it is Popperian to demand that a theory overcome a stiff observational hurdle. Only the Type I error rate is controlled in a statistical test, and a confirmation of the n + 1-th version is a falsification of the nth version; consequently, the error of false theoretical support must be considered to be the error associated with a false rejection of the null hypothesis. Hence, the theoretical prediction, fortified by a good-enough belt, must define the alternative hypothesis, and the null hypothesis must be its logical complement.

We emphasize that these same considerations apply to experiments in the hard sciences. Nature is just as unkind to physicists as it is to psychologists. If theory predicts a point value or function form, and if no good-enough region is used, then because all theories are false, sufficient precision will *always* result in a falsification. In both the hard and soft sciences, with sufficient precision and without a good-enough region, the results of experiments are always known in advance. It is only with the aid of a good-enough specification that one can avoid the paradoxical conclusions made inevitable by the prospect of perfect precision. So although the asymmetry between psychology and physics is indeed real, it is only real in the sense that the nature of the point values typically tested under the null hypotheses are different. In other words, although the mathematical sophistication of psychological theories does not yet permit the derivation of quantitative predictions, as is more typically the case in physics, the logical form of hypothesis testing is the same in that the substantive theory allows the logical derivation of the alternative and null hypotheses.

Let us now examine the directional null hypothesis. Under the good-enough principle, one must specify both direction and magnitude in advance of the experiment. If the statistical test indicates a possible increase less than that which is specified as good enough, the directional null hypothesis is retained. This would be true even when the direction is chosen at random. Thus, with perfect precision and a good-enough belt, the null hypothesis will *never* be rejected unless the magnitude of the effect is good enough. Without the good-enough

specification, random assignment of direction and perfect precision will result in a rejection half of the time. Thus, the good-enough principle does stiffen the observational hurdle in the case of a directional null hypothesis and perfect precision.

Admittedly, specifying good-enough values is difficult. As Walster and Cleary (1970) noted, "Regardless of the researcher's point of view, . . . he must make a judgement about the magnitude of effects of interest" (p. 248). The width of the good-enough belt depends on the state of the art of the theory and of the best measuring device available. It depends on the state of the art of the theory for two main reasons. First, a historical look at one's research program or an examination of a competing research program will help determine how accurately one's theory should predict in order that it be competitive with other theories. And second, in order to make predictions, one's theory may require parameters as input that themselves can only be predicted or measured to certain limits of accuracy. These limitations would need to be incorporated into the good-enough belt.

The width of the good-enough belt also depends on the best available measuring instrument. All "facts" are based on the theories of the measuring devices, theories that are false, so a certain amount of *systematic* error is liable to be present in any datum. The magnitude of this error will affect the width of the good-enough belt. *Random* error is not included in good-enough considerations, for such error is accounted for by the statistical hypothesis test.

Statistical Procedures

Over the last several years, researchers have shown increasing interest in testing hypotheses for and calculating confidence intervals about measures of effect size. For example, as Cohen (1990) wrote, "the primary product of a research inquiry is one or more measures of effect size. . . . You can attach a p value to it, but it is far more informative to provide a confidence interval" (p. 1310). Part of this emerging interest in effect sizes may be due to a broadening understanding that effect sizes are more indicative of substantive importance than is a p value, because the latter is strongly influenced by sample size. Another possible reason for the increased interest in effect sizes may be the relatively recent availability of meta-analytic techniques that often use effect sizes as data. A major reason for our interest is that our theories must make predictions that include a good-enough belt. These predictions are then used to specify statistical hypotheses that involve ranges, not point values, and range hypotheses concerning effect sizes can be tested using known statistical distributions. As Gigerenzer (chap. 11, this vol.) wrote, "We need rich theoretical frameworks that allow for specific predictions in the form of precise research hypotheses. The null hypothesis of zero difference (or zero correlation) is only one version of such a hypothesis—perhaps only rarely appropriate."

Our approach agrees with that ascribed by Gigerenzer to Neyman and Pearson. We state null and alternative hypotheses and discuss Type I error rate and power. We urge a thoughtful assignment of Type I error rate, and we feel that the assignment must be made in advance of the experiment. Each experimental result is one piece of the ongoing work in a research program.

Overall, we view the statistical test as if it were a scientific observational instrument, designed to ascertain whether or not a theoretical prediction is supported by "fact." The scientist states the prediction, includes a good-enough belt, and, knowing that the instrument has been designed to provide a certain resolution, specifies in advance that if the data support the theory, the instrument should be able to reveal the support. The whole world then peers into the instrument, and the theory either is or is not supported. The results are then made part of the historical record of the research program. If necessary, the positive heuristic is invoked to account for the anomaly. Regardless, the work of the research program continues, even if the program is replaced by a successful competing research program, until the replaced program is considered to be hopelessly degenerating.

Methods are available that are applicable to the problem of testing the range null hypotheses required by the good-enough principle. Hodges and Lehmann (1954) provided a procedure for showing in the one- and two-sample model that the magnitude of the effect exceeded a specified minimum value and for obtaining a confidence interval for the population effect. Hedges (1981) described methods for performing tests and obtaining confidence intervals for Glass' (1976) standardized effect size measure (the standardized effect size is the difference between two means expressed in standard deviation units). These were based on a normal approximation to the noncentral t distribution. Kraemer (1983) provided a central t approximation that could be used to the same end.

Seemingly, then, procedures are available that would allow the researcher to make decisions regarding the magnitude of an effect and, thereby, to test the range null hypotheses of the good-enough principle. Unfortunately, the Hodges and Lehmann (1954) test requires the use of their charts, from which accuracy is somewhat difficult to obtain, or a specialized computer program that is not available in the literature. In addition, the calculation of their confidence interval would also require a similar computer program, and as yet no one has solved the problem of showing that the effect was less than a specified magnitude.

Here we concentrate on test and confidence interval procedures involving Glass' standardized effect size, the correlation ratio in analysis of variance, and the square of the multiple correlation coefficient in regression analysis. There are two reasons for this specialization. First, the distributions of these effect size measures are known, and second, computer programs are available in the literature for performing the required calculations. We focus throughout on using the tests and confidence intervals for the purpose of determining when a confirming

instance has occurred, and we show that the approximations of Hedges and Kraemer are inadequate to this task under certain circumstances.

The One- and Two-Sample Cases

Various notation has been used to refer to the sample and population standardized mean difference in the two-sample case. For example, Hedges (1981) denoted these as g and δ, respectively, whereas Kraemer (1983) referred to them as d and δ. Some confusion can arise with either choice because, as discussed by these authors, the distribution of the sample standardized effect size is that of the noncentral t, whose noncentrality parameter is often denoted as δ. We adopt here Kraemer's notation, including using λ to refer to the noncentrality parameter of the noncentral t distribution.

In the one-sample case, the range null hypothesis will specify that the population mean will lie within (or outside of, depending on prediction) a specified range, here specified in population standard deviation units, of a predicted value. The sample statistic, d_1, is the difference between the sample mean and the specified value, expressed in sample standard deviation units. Similarly, in the two-sample case, the range null hypothesis will state that the population mean difference will lie within (or outside of, again depending on prediction) a range, again expressed in population standard deviation units, of a predicted value. Here, the sample statistic, d_2, expresses how far the difference in sample means is from the predicted value, expressed in sample standard deviation units.

Distribution Theory

In the one-sample case, let N observations be drawn from a normally distributed population with mean μ and variance σ^2. Then the sample mean $\bar{Y}$ will be normally distributed with mean μ and variance σ^2/N. Let the sample variance S^2 be an unbiased estimate of σ^2, and let the theoretically predicted value of the population mean be denoted μ_0. Then the theoretical prediction will state that $\delta_1 = (\mu - \mu_0)/\sigma$ is good enough. δ_1 is estimated in the sample by $d_1 = (\bar{Y} - \mu_0)/S$. Under the range null hypothesis, $t = \sqrt{N}\, d_1$ follows a noncentral t distribution with $N - 1$ degrees of freedom and noncentrality parameter $\lambda = \sqrt{N}\, \delta_1$.

In the two-sample case, let n_1 and n_2 observations be drawn from two normally distributed populations with means μ_1 and μ_2 and common variance σ^2. Then the sample mean difference $\bar{Y}_1 - \bar{Y}_2$ will be normally distributed with mean $\mu_1 - \mu_2$ and variance $N\sigma^2/n_1n_2$, where N is the total sample size. Let the pooled within-group sample variance S^2 be an unbiased estimate of σ^2, and let the theoretically predicted value of the population mean difference be denoted μ_0. Then the theoretical prediction will state that $\delta_2 = [(\mu_1 - \mu_2) - \mu_0]/\sigma$ is good enough. δ_2 is estimated in the sample by $d_2 = [(\bar{Y}_1 - \bar{Y}_2) - \mu_0)]/S$. Most often in

the two-sample case, psychologists are interested in estimating the population standardized mean difference from the sample standardized mean difference; if so, then $\mu_0 = 0$ in the formulas for δ_2 and d_2. Under the range null hypothesis, $t = \sqrt{n_1 n_2 / N}\, d_2$ follows a noncentral t distribution with $N - 2$ degrees of freedom and noncentrality parameter $\lambda = \sqrt{n_1 n_2 / N}\, \delta_2$.

Good-Enough Methodology

We describe the good-enough hypothesis-testing and confidence interval procedures in the two-sample design, first in the directional case where theory predicts that the effect will exceed a minimum. We then examine the method when a directional effect less than a maximum is predicted. We next describe the procedure for testing that the population standardized effect is within a good-enough range in either direction of prediction, and finally we describe the method for testing the prediction that the effect is outside a good-enough range in either direction of prediction. The one-sample case follows similar logic in all respects.

The first prediction to be examined, then, is that the effect is larger than Δ_S, the smallest magnitude of an effect, expressed in standard deviation units, that would be considered good enough. An effect smaller than this would constitute a falsifying instance. Then the null hypothesis would be

$$H_0\colon \delta_2 \leq \Delta_S.$$

This null hypothesis specifies a range for δ_2. Given the relationship between δ_2 and λ, the null hypothesis also specifies a range for the noncentrality parameter of the underlying noncentral t distribution of the sample d_2. Small values of the sample test statistic are consistent with the null hypothesis, thus only large values of the sample test statistic will lead to rejection of the null hypothesis. The critical value is chosen so that, under H_0, the sample test statistic will exceed the critical value no more than $100\alpha\%$ of the time. This Type I error rate depends on the true value of λ in the population.

Of course, there is only one true value of λ, and if it were known, the experiment would be unnecessary. The value of λ is unknown, thus the critical value for the test must be chosen so as to ensure that the Type I error rate does not exceed α for any value of λ allowed under the null hypothesis. The larger the true λ is, the greater the probability that the sample test statistic will be greater than any particular fixed value. Therefore, if the critical value is chosen so that the Type I error rate equals α when λ is at the limit allowed under H_0, then because all other values of λ under the null hypothesis are smaller than this upper limit, the Type I error rate under H_0 is guaranteed to be at most α. For the present directional null hypothesis, then, the critical value, denoted $CV_{1-\alpha}$, is chosen as the $100(1 - \alpha)$ percentile of the noncentral t distribution with $N - 2$ degrees of

freedom and with noncentrality parameter $\lambda = \sqrt{n_1 n_2/N}\,\Delta_S$. A computer program provided by Cooper (1968) and modified by Chou (1985) can be used to determine the critical value, or the normal approximation used by Hedges (1981) or the central t approximation used by Kraemer (1983) will yield an approximation to the critical value.

In order to find a confidence interval for δ_2, both Hedges (1981) and Kraemer (1983) solved for δ_2 in the equations approximating $CV_{1-\alpha}$. The results are approximate confidence intervals. An alternative would be to use the method described by Venables (1975) and the program of Cooper (1968) and Chou (1985) to solve for a confidence interval for λ, which then yields an exact confidence interval for δ_2. The procedure can be conceptualized in the following manner (see Kendall & Stuart, 1967, p. 206): Let θ denote the parameter for which the confidence interval is desired. Envision that a test of the hypothesis H_0: $\theta = \theta_0$ is to be performed, where θ_0 is one particular value of θ. Determine whether the observed sample statistic would fall into the acceptance region or the rejection region of the hypothesis test. Repeat this determination for all possible values of θ_0. If we aggregate the "acceptable" values of the parameter, we obtain the confidence interval.

For the hypothesis under consideration, the parameter of interest is the noncentrality parameter. Given the observed value of the test statistic, say t_{obs}, we ask for which values of λ_0 would we accept the null hypothesis H_0: $\lambda \leq \lambda_0$? This null hypothesis would be accepted for large values of λ_0 and rejected for small values of λ_0, thus the cutoff between these sets of λ values is that value of the noncentrality parameter for which t_{obs} equals the critical value of the test. If the value of λ that makes $CV_{1-\alpha}$ equal to the observed test statistic is denoted $\lambda_{1-\alpha}(t_{obs})$, then the confidence interval for λ is given by $\lambda \geq \lambda_{1-\alpha}(t_{obs})$. The confidence interval for δ_2 can be found from the confidence interval for λ by dividing by $\sqrt{n_1 n_2/N}$.

The next test to be examined is that of the prediction that the effect is smaller than Δ_M, the largest magnitude of an effect, expressed in standard deviation units, that would be considered good enough. Now the null hypothesis is

$$H_0\text{: } \delta_2 \geq \Delta_M.$$

Large values of the sample test statistic are consistent with the null hypothesis, so only small enough values of the sample test statistic will lead to rejection. The critical value, denoted CV_α, is chosen as the 100α percentile of the noncentral t distribution with $N - 2$ degrees of freedom and with noncentrality parameter $\lambda = \sqrt{n_1 n_2/N}\,\Delta_M$. If the value of λ that makes CV_α equal to the observed test statistic is denoted $\lambda_\alpha(t_{obs})$, then the confidence interval for λ is given by $\lambda \leq \lambda_\alpha(t_{obs})$. The confidence interval for δ_2 can be found from the confidence interval for λ by dividing by $\sqrt{n_1 n_2/N}$.

We next examine the procedure for testing the prediction that the population

standardized effect is within a good-enough range in either direction of prediction. This time the null hypothesis is

$$H_0\colon |\delta_2| \geq \Delta_M.$$

Large absolute values of the sample test statistic are consistent with the null hypothesis; consequently, only small absolute values of the sample test statistic will lead to rejection. Two critical values, denoted $CV_{\alpha/2}$ and $CV_{1-\alpha/2}$, could be chosen equal to the $100\alpha/2$ and the $100(1 - \alpha/2)$ percentiles of the noncentral t distribution with $N - 2$ degrees of freedom and with noncentrality parameter $\lambda = \sqrt{n_1 n_2/N}\,\Delta_M$. In the case of this hypothesis, however, it is perhaps easier to deal with the hypothesis expressed in terms of the square of δ_2 and the square of Δ_M, because then only one critical value derived from the noncentral F distribution is required. In addition, the procedure considered here can then be extended to other measures of association, presented later. The present null hypothesis is equivalent to

$$H_0\colon (\delta_2)^2 \geq (\Delta_M)^2.$$

Under this latter range null hypothesis, $F = (n_1 n_2/N)(d_2)^2$ follows a noncentral F distribution with 1 and $N - 2$ degrees of freedom and noncentrality parameter $\lambda = (n_1 n_2/N)(\delta_2)^2$. Large values of F are consistent with the null hypothesis, so the 100α percentile of the noncentral F distribution with 1 and $N - 2$ degrees of freedom and noncentrality parameter $\lambda = (n_1 n_2/N)(\Delta_M)^2$ is chosen as the critical value for the test, CV_α. If the value of λ that makes CV_α equal to the observed test statistic is denoted $\lambda_\alpha(F_{obs})$, then the confidence interval for λ is given by $\lambda \leq \lambda_\alpha(F_{obs})$. The confidence interval for the absolute value of δ_2 can be found from the confidence interval for λ by dividing by $n_1 n_2/N$ and taking the square root. A computer program written by Narula and Weistroffer (1986) can be used to find the critical value and the limit to the confidence interval.

Finally, the method for testing the prediction that the population effect is outside a good-enough range follows the same logic as the last procedure, with the upper tail probabilities of the noncentral F distribution substituted for the lower tail probabilities in the previous method. The null hypothesis

$$H_0\colon |\delta_2| \leq \Delta_S$$

is equivalent to

$$H_0\colon (\delta_2)^2 \leq (\Delta_S)^2.$$

The critical value, $CV_{1-\alpha}$, is equal to the $100(1 - \alpha)$ percentile of the F distribution with 1 and $N - 2$ degrees of freedom and noncentrality parameter $\lambda = (n_1 n_2/N)(\Delta_S)^2$. The confidence interval for λ is given as $\lambda \geq \lambda_{1-\alpha}(F_{obs})$, from which the confidence interval for the absolute value of δ_2 can be found by dividing by $n_1 n_2/N$ and taking the square root.

Power

For each of the good-enough tests, as for all standard null hypothesis tests, the power of the procedure depends on α, sample size, and how far the true population parameter is from that specified under the null hypothesis. For a given effect size specified under the null hypothesis and for a given sample size and Type I error rate α, a power curve can be drawn for the test by specifying effect size values under the alternative hypothesis and determining cumulative probabilities using the appropriate computer program.

Alternatively, one can specify the effect size under the null hypothesis, the Type I error rate α, and the desired power to detect a particular effect of interest under the alternative hypothesis and use the computer program to determine the required sample size. Note here that two effect sizes must be specified. In the cases in which it is desired to show that the effect size exceeds a particular value, we can indicate the nature of these two required effect sizes by using the terminology of Levin (personal communication, 1991): The limiting effect under the null hypothesis would be called the *maximum effect of noninterest*, whereas the effect specified for power calculations would be called the *minimum effect of interest*. As the true effect approaches the latter, matters get more and more interesting. Similarly, in cases in which it is desired to show that the effect size is less than a particular value, the terms become reversed: The limiting effect under the null hypothesis would be called the *minimum effect of noninterest*; and the effect specifed for power calculations would be called the *maximum effect of interest*.

Example

Let us illustrate the exact test and confidence interval procedures for a directional hypothesis and compare the confidence interval to the approximate intervals of Hedges (1981) and Kraemer (1983). Let us assume that theory-based considerations specify a directional good-enough prediction that the means should differ by at least 0.2 standard deviations, and let us also assume, as in an example provided by Kraemer (1983), that the sample sizes were $n_1 = n_2 = 10$. Based on these sample sizes and the hypothesized limit to the effect size, the hypothesized limit on the noncentrality parameter is calculated to be 0.4472. From the computer-generated cumulative percentiles of the noncentral t distribution with 18 degrees of freedom and noncentrality parameter 0.4472, $CV_{1-\alpha} = 2.229$. The power curve for this example is shown in Fig. 6.1. If, as in Kraemer's example, the observed effect size d_2 was 1.0, then the sample test statistic $t_{obs} = 2.236$. Hence, the directional null hypothesis is rejected, and it is concluded that a confirmation has occurred. In terms of a confidence interval, we find that for $\lambda_0 = 0.4538$, $CV_{1-\alpha}$ would equal $t_{obs} = 2.236$. Hence, the confidence interval for

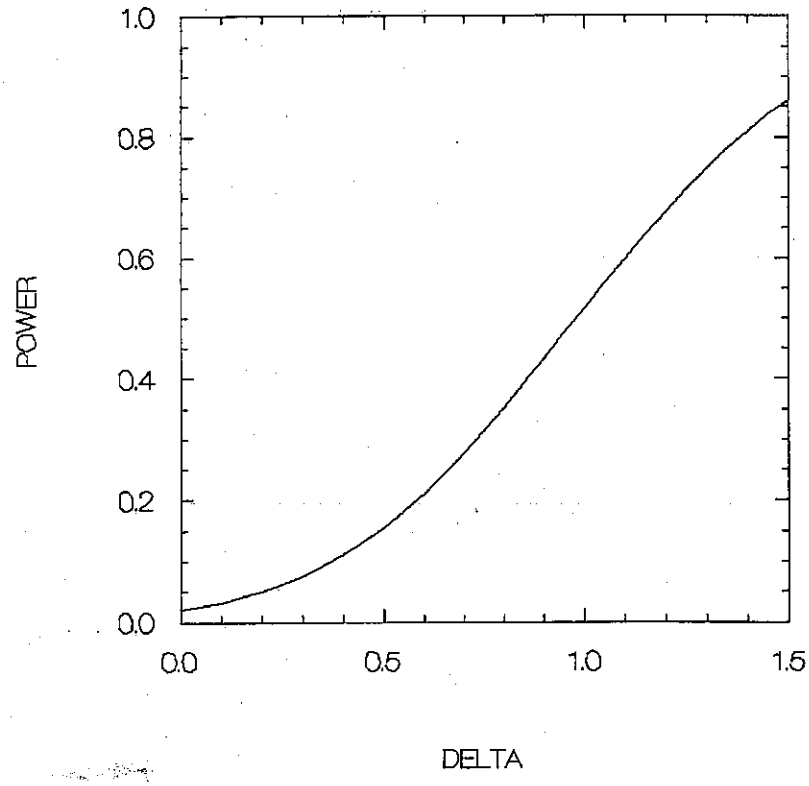

FIG. 6.1. Power for Kraemer example.

the noncentrality parameter is $\lambda \geq 0.4538$, from which is found the confidence interval for the effect size, $\delta_2 \geq 0.2029$. From Hedges's and Kraemer's respective approximations we would find for the confidence interval $\delta_2 \geq 0.1765$ or $\delta_2 \geq 0.2037$.

The adequacy of the approximations must be viewed from a Popperian perspective, recalling that the purpose of rigorous experimentation is to reveal possible confirming outcomes. If theory maintains that the effect size should be greater than 0.2 standard deviations, the Hedges approximation would not reveal the confirmation, whereas the Kraemer approximation and the exact approach would uncover it. In this sense, the Hedges approximation is conservative, here by about 13%. Unfortunately, although the Kraemer approximation typically yields a value close to the exact limit, it is slightly liberal. Whereas the Hedges approximation would miss some confirmations, the Kraemer approximation would typically "discover" too many. For these reasons, it is best to use the exact method.

Comparison with Cohen's (1988) Method of "Proving the Null Hypothesis"

In using the present methodology to confirm the prediction that an effect is at most some particular magnitude, we are able to show that an outcome is effectively zero. Elaborating on this application may help to highlight the distinctive features of the good-enough methodology, especially when this methodology is compared to that suggested by Cohen (1988, pp. 16–17) for the problem often referred to as "proving the null hypothesis." It is well known that a nonrejection of the standard null hypothesis does not allow the same kind of conclusion to be drawn as would a rejection. If one wishes to confirm the prediction that an effect is trivial, then the logical complement must be set up as the null hypothesis. Cohen (1990), on the other hand, maintained that one could use the traditional null hypothesis testing methodology probabilistically to "prove the intended null hypothesis of no more than a trivially small effect" (p. 1309).

In Cohen's (1988) view, the investigator may validly conclude that an effect is trivial under certain circumstances. The experiment should be set up as if testing the standard null hypothesis, with a specified Type I error rate, which he called a. The sample size should be chosen so that the power of a standard null hypothesis test, which he called $(1 - b)$, is high for the detection of a specified *trivial* effect size, which he called i (for iota). Conceptually, if the power is very high and the null hypothesis is *still* not rejected, then the effect size is very likely to be less than i. "Thus," wrote Cohen (1988), "in using the same logic as that with which we reject the null hypothesis with risk equal to a, the null hypothesis can be accepted in preference to that which holds that [effect size] $= i$ with risk equal to b. Since i is negligible, the conclusion that the population [effect size] is not as large as i is equivalent to concluding that there is 'no' (nontrivial) effect" (p. 16).

There are certain difficulties with the Cohen approach. First, because Cohen indicated (1988, p. 104) that the null hypothesis for his procedure has been mitigated "to mean 'trivially small,'" the conclusion that the effect is trivially small is drawn from a nonrejection of the null hypothesis, rather than from a rejection of its logical complement. This seems to be at odds with standard statistical logic. In addition, some confusion results from using the same notation and terminology for this technique and for the standard null hypothesis test. If the risk in his procedure is now b, the power of the test cannot be $(1 - b)$, and it is not clear just what the role of a is. As suggested by Gigerenzer (chap. 11, this vol.), we should "point out the confused logic of the hybrid, and insist on consistency." The good-enough methodology allows us to do so.

As we described earlier, good-enough methodology provides a test that allows the conclusion that an effect is trivially small. In order to draw this conclusion, say in a two-sample test, the null hypothesis must be

$$H_0\colon (\delta_2)^2 \geq i^2,$$

so that the alternative hypothesis is

$$H_1: (\delta_2)^2 < i^2,$$

where we use i as the criterion for triviality, as in Cohen's procedure. The critical value for this test is the 100α percentile of the noncentral F distribution.

We can discuss the power and Type I error rate for this procedure in terms of Cohen's a and b (defined for his standard null hypothesis test). The value b is the probability that a sample F ratio will be less than Cohen's critical value (resulting in nonrejection of the standard test). Cohen's b is found by equating the Cohen critical value to the $100b$ percentile of the noncentral F distribution, with the noncentrality parameter based on the effect size being equal to i; but this is exactly how the critical value for the good-enough procedure is found. Hence, Cohen's b is equivalent to the Type I error rate, α, for our procedure. It is for this reason that Cohen called b the risk in his procedure. Thus, if the triviality cutoffs are the same in the two procedures, and if b is set equal to α, the critical values for the procedures are the same. Finally, because $(1 - a)$ is the probability of accepting Cohen's null hypothesis when the noncentrality parameter is zero, it is also the power of the good-enough test for that condition; that is, $(1 - a)$ is the maximum power of the test that allows the conclusion that the true effect is less than i. As described earlier, a power curve can be drawn for the test. Figure 6.2 shows the power curve for an example from Cohen (1988, p. 58) in which $a = .05$ and $b = .05$. Then the Type I error rate $\alpha = b = .05$ and the maximum power is $(1 - a)$, or 0.95, for detecting an effect size less than $i = 0.20$. Note that the power for the procedure is adequate only if the true effect were less than 0.06.

Other Applications

The methods based on the noncentral F distribution can be extended to the analysis of variance model, in which case the good-enough limits Δ_S and Δ_M would be expressed in terms of the population correlation ratio η^2 (the proportion of total variability explained by group membership) and to the regression model, in which case the good-enough limits would be expressed in terms of the population squared multiple correlation coefficient R^2. In fixed-effect analysis of variance, under a range null hypothesis the sample F statistic follows a noncentral F distribution with noncentrality parameter $\lambda = N\eta^2/(1 - \eta^2)$ (see Timm, 1975, p. 365), whereas if the values of the predictors in regression analysis are assumed fixed, the sample F statistic follows a noncentral F distribution with noncentrality parameter $\lambda = NR^2/(1 - R^2)$ (see Anderson, 1958, p. 93). Given these relationships between the good-enough limits and the noncentrality parameters, the methods described in the nondirectional two-sample case can be used to test hypotheses and calculate confidence intervals for the parameters of interest.

As an example, assume a researcher theorizes that three measures of mood

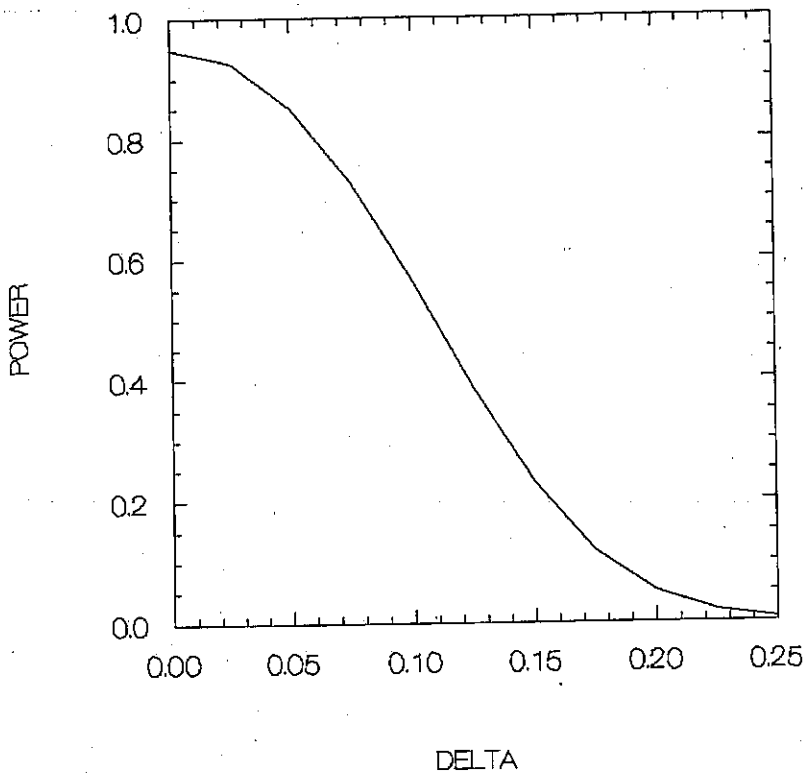

FIG. 6.2. Power for Cohen example.

should not be related to a cancer patient's perception of pain. It is felt that the mood variables should be able to account for at most $\Delta_M = 25\%$ of the total variability in pain scores for the relationship to be considered substantially trivial. Hence, the null hypothesis would be

$$H_0: R^2 \geq .25.$$

Assume a sample of $N = 90$ patients is observed and the sample squared multiple correlation coefficient was equal to .15. In terms of the noncentrality parameter, the null hypothesis can be written

$$H_0: \lambda = NR^2/(1 - R^2) \geq 90(.25)/(1 - .25) = 30.$$

The critical value, 5.2837, is found from the computer program as the fifth percentile of the noncentral F distribution with 3 and 86 degrees of freedom and noncentrality parameter equal to 30. Because the F statistic of 5.0588 is smaller than the critical value, a confirmation of the theory has occurred. The confidence

interval is given by $R^2 \leq .24$. Note that the standard F test would yield significance, with $p < .01$, allowing the conclusion that the population R^2 is nonzero, but that is of no interest in terms of theory confirmation and falsification.

PSYCHOLOGY AND PHYSICS RECONSIDERED

We have argued that significance-testing in psychology, when fortified with the good-enough principle, is not defective when compared with physics, and the methodological asymmetry noted by Meehl (1967) is more apparent than real. We would now like to reconsider the relationship between psychology and physics in this light. It would appear that nothing is so revealing of our inadequacies as a discipline than to compare our methodologies and achievements with that of physics. This is indeed a favorite strategem of those who are critical of the scientific pretensions of psychological research. In our view, however, this strategem is effective only when one adopts a view of physics that is overly optimistic and idealized, and one that is neglectful of the actual record of scientific practice in this discipline. Fortunately, more realistic appraisals of how to relate the two disciplines are beginning to emerge (e.g., D'Andrade, 1986). In a fascinating study, for example, Hedges (1987) was able to compare the empirical cumulativeness of research in both physics and psychology. By examining 13 quantitative reviews of the particle physics literature, and quantitative reviews of six domains of social science research, Hedges (1987) was able to conclude that psychological research is not disadvantaged in the cumulativeness of its findings when compared to this elite branch of physics. According to Hedges (1987),

> What is surprising is that the research results in the physical sciences are not markedly more consistent than those in the social sciences. The notion that experiments in physics produce strikingly consistent (empirically cumulative) results is simply not supported by the data. Similarly, the notion that experiments in the social sciences produce relatively inconsistent (empirically noncumulative) results is not supported by the data either. (p. 450)

Indeed, the rate of inconsistency and disagreement exhibited by the reviews was about 45% in either discipline, when no studies were eliminated from the reviews. (Consistency of findings improves in both disciplines when studies are omitted.) It would appear, then, that not only is there considerable similarity in hypothesis-testing methodologies between the two disciplines, as argued in the previous section, there is also considerable similarity evident in the empirical cumulativeness of research findings and in the quantitative methodologies used to detect it in the various literatures.

As Hedges (1987) noted, however, in the absence of similar kinds of studies, any firm generalization about the relative empirical cumulativeness of physics

and psychology must be limited to the research domains sampled. But Hedges (1987) was able to show just what might result when one excavates and reconstructs a literature. At least with respect to the research domains sampled here, there is little reason for psychologists to bemoan their supposed unhappy lot, and little reason as well to idealize the consistency of findings in particle physics. One additional case study is reported to support this view. The point of this exercise is not, of course, to disparage the workings and achievements of physicists, but rather to point out that scientific activity, in either psychology or physics, is a difficult and often messy affair. Both disciplines have their own problems, to be sure, but matters are not improved and no useful end is served when we engage in idolatrous veneration of physics at our own expense.

The Problem of Solar Neutrinos

Pinch (1985) provided a case study of theory testing in physics that is illustrative of many of the issues that we have touched on, such as the requirements of good-enough hypothesis testing, the inconclusiveness of refutations, the necessity for operating with conventional assumptions, and the difficulty in testing auxiliaries and *ceteris paribus*. The *problem of solar neutrinos* emerged during the course of putting the nuclear astrophysical theory to a test. The theory attempts to account for the evolution and structure of stars. Its central tenet purports that stars burn hydrogen. Although this tenet had been widely accepted within the astrophysical research community, it proved difficult to subject to a severe test, at least until 1958. At that time it was suggested that one way to test the "hydrogen-burning" hypothesis was to construct an apparatus that made it possible to detect solar neutrinos, which are a by-product of hydrogen burning. This was a daunting challenge, because neutrinos are exceedingly difficult to detect. One method involved the use of a large (100,000 gallon) tank of dry-cleaning fluid that was sunk one mile into the earth in an abandoned mine shaft. The construction of the apparatus was begun in 1964 but was not completed until 1967. Exceedingly complex mathematical calculations, aided by numerous simplifying assumptions, initially determined that the astrophysical theory would be corroborated if the detection methodology captured 40 $\pm$ 20 SNUs (Solar Neutrino Units). Before the experiment was actually conducted, however, the prediction was downsized to 19 $\pm$ 11 SNUs. Clearly, "deriving" a prediction from the theory was quite difficult, and involved not a "point" prediction per se but a point prediction within a specified "good-enough" range. Indeed, according to Pinch (1985), "The prediction of a solar neutrino capture rate depends on a whole morass of theories and experimental data. These are not even drawn from one single area of science" (p. 175). Further, as Pinch (1985) told it, input parameters had to be juggled in order to provide a realistic model of the Sun. "It is difficult to see," wrote Pinch (1985, p. 176), "how the derivation of the final prediction could be described as a process of deduction."

By August 1967 the results of the study were available, and they did not appear to support the (good-enough) prediction of the theory. Indeed, so low was the signal that the result could only be reported at an upper limit of 6 SNUs, a value that is compatible with background radiation. A second experimental run, aided by technical improvements, reported an even lower rate of neutrino capture (upper limit of three SNUs). By all appearances, then, the astrophysical theory sustained a stunning refutation as a result of this "crucial experiment."

In spite of the putative refutation, however, the theory was not abandoned. Indeed, the hard core assumption that stars burn hydrogen was never in doubt. The refutation only invited appeals to *ceteris paribus*. According to Pinch (1985), "It should be emphasized that none of the revisions were to the basic theory itself, that remained largely unaltered. Rather it was amongst the large number of input parameters necessary for the calculation that the adjustments were made" (p. 180). New calculations of SNU capture rate, plus other recent findings, led to a new post hoc "prediction" of 7.5 ± 3 SNUs. Although the observed capture rate was still out of this good-enough range, it was thought that other refinements would bring the predicted SNU capture rate into line with what was actually observed. It is true to say, of course, that different members of a research community will evaluate the data and what is at stake theoretically in different ways. Critics of the astrophysical theory could see a refutation if strong doubts are expressed about the legitimacy of initial conditions, methodological and simplifying assumptions, parameter estimation, and the like. As Pinch (1985) pointed out:

> Initial conditions, like all basic statements, are themselves only accepted by conventional processes. There is nothing to stop scientists, if they so wish, from refusing to accept the new initial conditions. This means that they need not be bound by any apparent consistency or contradiction between theory and observation statement. By refusing to accept the initial conditions, they, in effect, refuse to accept the outcomes of the test of theory based on those conditions. The conventional element at the heart of scientific practice allows for the possibility that agreement over consistency or contradictions cannot be reached during a crucial test. (p. 181)

This shows clearly, of course, that researchers are not bound by the tyranny of data. The failure of a prediction, held out on the eve of an experiment, does not automatically falsify a theory. Theories can be "juggled" so that support can be claimed if data merely satisfies, in some good-enough way, and sometimes after the fact, some anticipated consequence of the theory. But one can only claim support; it can never be compelled from the research community.

The solar neutrino example illustrates a number of themes. It shows, first of all, that the derivation of "point" predictions is a difficult affair, involving the juggling of parameters, simplifying assumptions, the uncertainty of derivations through the morass of auxiliary theories, and the specification of a good-enough

range. It also shows that corroboration and refutation are only uncertainly and conventionally linked to the success or failure of predictions. Thirdly, the neutrino example illustrates how theories are clung to tenaciously and the appeal procedure by which theorists protect the hard core of the theory from refutation. Certainly, to the naive falsificationist, the solar neutrino example represents a serious departure from acceptable scientific method. Yet this brand of falsificationism does not credibly reconstruct the actual practice of working scientists. There is a lesson to be drawn for research psychologists: The difficulties of deriving testable predictions through the morass of uncertain auxiliary theories, the uncertainty involved in the estimation of parameters, the stubborn refusal to allow theories to be refuted, even to the point of pursuing after-the-fact strategems, should not be used as an indictment against the research practices of psychologists. All of this is well in evidence in astrophysics, and may well constitute standard practice for working scientists, whatever the discipline.

CONCLUSION

The purpose of this chapter was to respond to two popular criticisms of psychological research. It is claimed, for example, that theory appraisal in psychology is an impoverished affair, due to defeciencies in the use of significance testing. The criticism is then thought to be driven home by comparing the research practices of psychologists with what is taken to be the standard case in physics. In contrast, we have argued that hypothesis-testing in psychology, when fortified with the good-enough principle, is not rationally disadvantaged when compared against hypothesis testing in physics. We also argued that slow progress and lack of cumulative findings are not the indictments of psychology they are often made out to be. We showed, for example, by appealing to Hedges (1987), that the *empirical* cumulativeness of at least some domains of psychology is strikingly comparable to what is the case in particle physics. We also argued that evidence for *theoretical* cumulativeness is best sought by reconstructing the various literatures in light of the methodology of research programs as described by Lakatos (1978a). We complained about the tendency of critics to unfairly hold up an idealized vision of research practices in physics as a standard by which to indict psychological practices, and offered two case studies that should encourage not only a more realistic appraisal of what is actually the case in physics, but also a more charitable appraisal of what is the case in psychology.

Meehl (1990) suggested that one way to improve the prospects of psychological research was to increase the mathematical proficiency of our graduate students. This is an interesting proposal that merits serious consideration. We would like to add our own suggestion. In our view it is not simply enough to train graduate students to be conversant with various theories, and to be skilled users of statistical methodologies. What must also be inculcated is the view that

theories are part of research programs that must be evaluated historically in light of certain criteria that help us appraise growth, progress, and cumulativeness. In other words, research programs must be rationally reconstructed. Meehl was more right than not when he complained about the amount of "naive guessing" that goes on in psychological research, much of which is embarassingly atheoretical. Many readers of this volume undoubtedly have sat on dissertation committees where the chief rationale for doing a particular study is sheer novelty, and absent any larger concern for how the anticipated results might constitute growth in a research program. Typically the literature review goes back only a handful of years, or else touches on only that part of the relevant literature that addresses some local concern of interest. It is not enough to simply "know" a literature. One must also reconstruct it in such a way that it can be seen to reveal the cumulative character of the enterprise. The methodology of research programs provides an impressive armamentum of tactics to this end. The general procedure, according to Lakatos (1978a), is to first provide a rational reconstruction: "One tries to compare this rational reconstruction with actual history, and try to criticize both one's rational reconstruction for lack of historicity and the actual history for lack of rationality" (p. 53). Whether or not one finds the Lakatosian framework congenial, and it is not without its critics, the main point is that we will never find cumulative progress in our research until we look for it. And rational appraisal of psychological research, and our appreciation of its scientific character, will always hinge on how we appraise growth in knowledge.

REFERENCES

Anderson, T. W. (1958). *An introduction to multivariate statistical analysis.* New York: Wiley.

Bolles, R. (1962). The difference between statistical hypotheses and scientific hypotheses. *Psychological Reports, 11,* 639–645.

Case, R. (1985). *Intellectual development: Birth to adulthood.* Orlando, FL: Academic Press.

Chou, Y. (1985). A remark on algorithm AS5: The integral of the noncentral *t*-distribution. *Applied Statistics, 34,* 102.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45,* 1304–1312.

Cooper, B. H. (1968). Algorithm AS5: The integral of the noncentral *t*-distribution. *Applied Statistics, 17,* 193–194.

Cooper, H. (1989). *Integrating research: A guide for literature reviews.* Newbury Park, CA: Sage Publications.

D'Andrade, R. (1986). Three scientific world views and the covering law model. In D. Fiske & R. Shweder (Eds.), *Metatheory in social science: Pluralisms and subjectivities* (pp. 19–41). Chicago: University of Chicago Press.

Feldman, K. (1971). Using the work of others: Some observations on reviewing and integrating. *Sociology of Education, 44,* 86–102.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5,* 3–8.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6,* 107–128.

Hedges, L. (1987). How hard is hard science, how soft is soft science? The empirical cumulativeness of research. *American Psychologist, 42,* 443–455.

Hedges, L., & Olkin, I. (1985). *Statistical methods for meta-analysis.* Orlando, FL: Academic Press.

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

Jackson, G. (1980). Methods for integrative reviews. *Review of Educational Research, 50,* 438–460.

Kendall, M., & Stuart, A. (1967). *The advanced theory of statistics* (Vol. 2). New York: Hafner.

Kraemer, H. C. (1983). Theory of estimation and testing of effect sizes: Use in meta-analysis. *Journal of Educational Statistics, 8,* 93–101.

Ladas, H. (1980). Summarizing research: A case study. *Review of Educational Research, 50,* 597–624.

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, pp. 8–101). 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, pp. 128–210). Cambridge, England: Cambridge University Press.

Lakatos, I. (1978c). Popper on demarcation and induction. In J. Worrall & G. Currie (Eds.), *The methodology of scientific research programs: Imre Lakatos philosophical papers* (Vol. 1, pp. 139–167). Cambridge, England: Cambridge University Press.

Lakatos, I. (1978d). Introduction: Science and pseudoscience. In J. Worrall & G. Currie (Eds.), *The methodology of scientific research programs: Imre Lakatos philosophical papers* (Vol. 1, pp. 1–7). Cambridge, England: Cambridge University Press.

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

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.

Meehl, P. (1986). What social scientists don't understand. In D. Fiske & R. Shweder (Eds.), *Metatheory in social science: Pluralisms and subjectivities* (pp. 315–329). Chicago: University of Chicago Press.

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

Morrison, D., & Henkel, R. (Eds.). (1970). *The significance test controversy.* Chicago: Aldine.

Narula, S., & Weistroffer, H. (1986). Computation of probability and non-centrality parameter of a non-central F-distribution. *Communications in Statistics B, 15,* 871–878.

Phillips, D. C., & Nicolayev, J. (1978). Kohlbergian moral development: A progressing or degenerating research program? *Educational Theory, 28,* 286–301.

Pinch, R. (1985). Theory testing in science—the case of solar neutrinos: Do crucial experiments test theories or theorists? *Philosophy of Social Science, 15,* 167–187.

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

Popper, K. (1963). *Conjectures and refutations.* London: Routledge & Kegan Paul.

Rosnow, R., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 44,* 1276–1284.

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

Serlin, R. C., & Lapsley, D. K. (1990). Meehl on theory appraisal. *Psychological Inquiry, 1*, 169–172.

Spielman, S. (1974). The logic of tests of significance. *Philosophy of Science, 41*, 211–226.

Timm, N. (1975). *Multivariate analysis*. Monterey, CA: Brooks/Cole.

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.

Venables, W. (1975). Calculation of confidence intervals for noncentrality parameters. *Journal of the Royal Statistical Society (B), 37*, 406–412.

Walster, G. W., & Cleary, T. A. (1970). Statistical significance as a decision rule. In E. Borgatta & C. Bohrnstedt (Eds.), *Sociological Methodology* (pp. 246–254). San Francisco: Jossey-Bass.