Journal of Methods and Measurement in the Social Sciences,
Vol.12, No. 1, 5-36, 2021

What Behavioral Scientists Are Unwilling to Accept

Lewis Petrinovich
University of California, Riverside

Meehl (1986) published a brilliant paper with the title "What Social Scientists Don't Understand." I believe that that a better characterization of the situation is to broaden the reference class to include behavioral scientists in general. Also, I believe that most behavioral scientists do understand these issues at some level, but that they are not willing to accept the implications that an explicit understanding would force.

Concern has been expressed, for the past 40 years or so, that the strategies of research design and statistical analysis used by behavioral scientists are woefully inadequate to support a progressive scientific enterprise. In this article I will summarize the nature of some of these concerns, and will identify some of the impediments that they impose to the development of progressive conceptual frameworks adequate to the task of achieving an understanding of the behavior of complex organisms in their environment. Although there has been little disagreement regarding the soundness of the methodological criticisms that have been made, there is little reason to believe that the methodological and statistical criticisms have had any great impact on the activities of either journal editors or research scientists. A review of these concerns is appropriate at this time because a number of articles have appeared recently that defend the *status quo*, and which are based either on faulty premises or a questionable view of the problems that impede scientific progress. And, hopefully, yet another critique might enhance the rate at which we arrive at a more satisfying state of affairs.

Following this polemic I will suggest some orienting attitudes, research procedures, and analytic strategies that should lead us to a better understanding of the universe of events we, as behavioral scientists, are attempting to understand. These alternative views are based on current developments in the philosophy of science and entail the use of design and analytic strategies that are of sufficient complexity to permit advances in the understanding of the behavior of organisms in their environment.

Context of Discovery

Although the problem of demarcation, differentiating between science and non-science, is beyond the scope of this paper, I will consider the distinction between the contexts of discovery and justification, a distinction that is often discussed when considering the demarcation problem (e.g. Popper, 1959). A consideration of these contexts will be useful because the crux of much of my argument hinges on the assumption that we need to adopt strategies and methods that have high heuristic value, rather than those that merely speak to issues of justification.

Popper (1959, p. 31) writes, "The initial stage, the act of conceiving or inventing a theory, seems to me neither to call for logical analysis nor to be susceptible of it." He believes there is no such thing as a logical method of having new ideas, or, even, of reconstructing the process, because every discovery contains an irrational element. Although he considers intuitions to be important to the development of science they are consigned to the context of discovery, and placed outside the realm of empirical science. He concentrates his analysis on the context of justification (which he considers to be subject to rational analysis), and adopts a criterion of falsification: for a statement to be scientific it must lead to formulations relating to concrete observations, some of which would conflict with theoretical expectations.

This line of argument led Feyerabend (1975) to the more extreme position that, in reality, we are dealing with a single uniform domain of procedures (which contains an element of irrationality throughout), which contains elements each of which are important for the growth of science. Feyerabend considers the development of science to proceed by the invention and articulation of alternative hypotheses (as Popper emphasizes), but comes to the conclusion that the distinction between Popper's contexts of discovery and justification breaks down because, "What remains are aesthetic judgements, judgements of taste, metaphysical prejudices, religious desires, in short, *what remains are our subjective wishes….*" (1975, p. 285).

Context of Justification

I suggest that the demarcation criterion, establishing the line between science and non-science, depends on the terms that define the adequacy of observation in terms of reliability, and that the primary concern of observational procedures that are exclusively scientific should be on justification to a great extent. However, procedures that enhance theoretical discovery should continue to be emphasized forever: the identity and nature of the important variables to be examined that will permit an understanding of the phenomena of interest are always of paramount importance, and discovering their identity imposes immense difficulties. Clearly, as Tukey (1969, p. 83) has written, "To concentrate on confirmation, to the exclusion or submergence of exploration, is an obvious mistake." The context of discovery is contributed to by anyone who considers a problem (philosophers, novelists, musicians, poets, priests, or scientists), and some of this endeavor can well have a reasoned logic. However, only when measurement begins and the context of justification is involved does the work uniquely identified as science begin. Thus, we must never lose sight of the problems involved in discovery, and must stress continually the heuristic value of scientific theories.

H. I. Brown (1977, p. 10) argues in much the same manner; "Perhaps the most important theme of the new philosophy of science is its emphasis on continuing research, rather than accepted results, as the core of science. As a result, analysis of the logical structure of completed theories is of much less interest than attempting to understand the rational basis of scientific discovery and theory change."

The Importance of Alternative Hypotheses

One of the most important ways to enhance the likelihood of discovering the strengths and limitations of theories and to lead to the discovery of needed information is to always frame empirical tests of hypotheses in terms of possible alternatives. By developing sets of alternatives we avoid some of the limitations on the strength of generalizations produced due to a failure to consider problems that occur with the use of a single guiding hypothesis. One of the earliest statements regarding the importance of considering alternative hypotheses when engaging in scientific research was made by the geologist, T. C. Chamberlin (1897). He urged scientists to use what he called the method of multiple working hypotheses, because "Each hypothesis suggests its own criteria, its own means of proof, its own method of developing the truth; and if a group of hypotheses encompass the subject on all sides, the total outcome of means and of methods is full and rich." (p. 846) One problem he identified is that when one used only a single hypothesis "There springs up also unwittingly a pressing of the theory to make it fit the facts and a pressing of the facts to make them fit the theory." (p. 840) Rather than employing a single hypothesis "...the effort [should be] to bring up into view every rational explanation of the phenomenon in hand and to develop every tenable hypothesis relative to its nature, cause or origin, and to give to all of these as impartially as possible a due place in the investigation." (p. 843).

The physicist, J. R. Platt (1964) endorsed Chamberlin's views and offered his version which he called "Strong Inference." He argued that one should apply the following steps to every problem in science: (1) devise alternative hypotheses; (2) devise a crucial experiment (or several of them) with alternative possible outcomes, each of which will, as nearly as possible, exclude one or more of the hypotheses; (3) carry out the experiment and evaluate the results; (4) recycle the procedure, making subhypotheses to refine the possibilities that remain, and investigate any new possibilities revealed by the experiment.

Platt points out that steps 1 and 2, outlined above, require intellectual inventions, which must be cleverly chosen. A view that emphasizes the context of discovery as being of paramount importance to the development of science rather than placing the major emphasis on the context of justification. Although justification is considered to be important the problems involved in justification are relatively simple compared to those involved in discovery.

Alcock (1989) clearly illustrates the power that results from casting alternative hypotheses when he considers the reasons for male infanticide in hanuman langur monkeys of India. It has been observed that, when an adult male displaces a dominant male who has a harem of females, the usurper kills the existing infants in the harem. Upon the death of her nursing infant the mother becomes sexually receptive at once and will copulate with, and be fertilized by, the new male harem master. In this way the infanticidal male increases the number of descendants he produces (because the female otherwise would not ovulate for 2-3 years), and more of his genes will be transmitted to the next generation.

Alcock considers this sexual competition explanation to be plausible, but evaluating it alone would provide only weak evidence for its plausibility. Rather, one should compile a list of alternative hypotheses for infanticide by male langurs, all of which would predict infanticide, but which would have other, non-overlapping predictions that, ideally, could be used to eliminate all but one of the alternative hypotheses. He developed three hypotheses that have at some time been proposed: (1) the above sexual competition hypothesis; (2) a cannibalism hypothesis assuming that, by consuming the infant, the male gains a high-protein meal which allows him to survive the rigors of the takeover effort more successfully; (3) a social pathology hypothesis which does not consider infanticide to be adaptive at all, but as a social pathology induced by artificial provisioning and the resulting overcrowding.

He derives four different predictions that bear on the hypotheses: some of the predictions are not made by one or more of the hypotheses, some are predicted (given certain special conditions), and others are critical, to the extent that if they do not occur the hypothesis is eliminated. Alcock is able, on the basis of the data related to the different predictions, to cast such severe doubt on the social pathology hypothesis that it is almost eliminated from further consideration. The likelihood of the cannibalism hypothesis is weakened considerably by the evidence, while the sexual competition hypothesis remains the most plausible in light of the predicted outcomes (see Alcock 1989, pp. 14-19 for the arguments).

The adequacy of the specific arguments are not important in the present context. What is important is that by devising alternative hypotheses, the likelihood of each of the three hypotheses offered to account for infanticide could be assessed through the process of clarifying the types of evidence that would bear on each hypothesis, indicating what evidence would be needed to evaluate each hypothesis, examining that evidence, and reflecting the obtained data on each hypothesis. The whole process can now be repeated, making additional predictions, incorporating other plausible hypotheses that might be appropriate, and considering all of the hypotheses in the light of yet additional evidence that might be discovered.

It is not only of value to consider alternative hypotheses when investigating a research area at a single level of discourse. As Cronbach (1986) has pointed out, even greater value might result when we consider our findings "...in combination with other reports and with beliefs from other sources..." (p. 95). He also suggests (p. 98), that "We can rarely see a topic in proper perspective if our inquiry employs resources from only one discipline." Ghiselin (1971) has pointed out that just such a value was realized in the biological sphere when both Darwin and Wallace read the economist Malthus, and were led independently to formulate the theory of natural selection; a case where established, normal theory in one discipline (economics) led to revolutionary theory in another (biology).

It might well be the case that not only should resources from other disciplines be considered, but that findings at related levels of analysis should be incorporated as well. If we wish to move toward the realist's goal of developing theories that "cut nature closer to the joints," the findings from disciplines at different levels of analysis (e.g. social, behavioral, physiological, neurophysiological) should be scrutinized as carefully as the alternative possible hypotheses at the same level. This does not imply that any level will be reduced ontologically to another, but only that the principles and processes operating at one level might provide insights regarding those at another, and that boundary conditions can be set on any one level by principles and processes operating at other levels.

A good example of how principles at one level influence those at another is afforded by the study of visual perception (Hochberg, 1988). Major advances in our understanding of neurophysiological mechanisms involved in perception have been forced by a consideration of phenomenological data. Physiological views about such things as lateral inhibition, hue, and contrast were designed to fit the phenomenological facts of perception, and the proposed physiological mechanisms were conceived to make direct comparisons and calculations based on the relations between different aspects of the proximal stimulation. Hochberg (1988, p. 232) summarizes this view as follows, "But the historical facts seem clear: *phenomenology has predicted more of recent neurophysiology than vice versa, and indeed if we wish eventually to be able to predict perceptual experience, then some explicit relationship between appearances and physiology must be provided*."

Hebb's successful development of a heuristically valuable theory (Hebb, 1949) at the level of the neurophysiological and psychological mechanisms involved in learning and perception was due to the fact that he started with phenomenology, considered known neurophysiological mechanisms, speculated about the probable relationships, and that he and his colleagues revised his theory on the basis of advances in neurophysiology (Goddard, 1980; e.g. Milner, 1957). The study of the physiology of visual perception, as well as the power of Hebb's theoretical formulations, can be counted as scientific success stories. I attribute much of this success to be due to recognition of the existence of factual constraints at interdisciplinary levels. The importance of the existence of such constraints became apparent through interdisciplinary considerations. Another way of stating this point is that, rather than achieving a molecular, ontological reduction of one discipline to another, it is possible to achieve a molar, theoretical reduction through the identification of common functional principles, as exemplified by the triumphs of the Darwinian synthesis (see Petrinovich, 1976).

Multitrait-Multimethod Procedures

The importance of devising alternative hypotheses, especially when attempting to evaluate explanations of the complex behavior of organisms interacting with one another in an ecologically representative context, is quite apparent. However, not only is the method of alternative hypotheses seldom used in behavioral science, there exists an additional problem. Whenever one is going to investigate a concept there must be an adequate operational translation of the theoretical constructs involved if we are to relate theoretical propositions to observables, and different theoretical propositions to one another.

To achieve an adequate characterization of a concept it becomes essential to use the procedures recommended to establish construct validity (Cronbach & Meehl, 1955), and to employ the multitrait-multimethod (MTMM) approach developed by Campbell & Fiske (1959), and used so compellingly by Hammond, Hamm, and Grassia (1986). The discussion regarding traits that will be presented in this section can be broadened to include any concept or theoretical statement. The procedures to establish an underlying trait based on such things as item responses are the same as those involved in establishing a concept on the basis of observed behavior. According to these views the basic considerations underlying all of our problems concern adequate measurement procedures. If we are ever to realize a progressive science, our efforts must be based on sound measurement procedures. Elsewhere I have argued (Petrinovich, 1979) that it is important to develop and evaluate theories by utilizing the methods of construct validity in the same manner as when one establishes the construct validity of a test. The central idea is that a test (or observation) involves the measurement of some latent attribute or quality that is not operationally defined: it is a postulated attribute assumed to be reflected in test performance (or observed behavior). The intended focus is on trait quality rather than test behavior, scores on various criteria, or on results dependent on the method that is used. It is essential, when seeking to establish construct validity, to gather evidence from several disparate sources, including different traits, and different methods of measurement. The numerical statement of the construct validity of a purported indicator of a latent construct is the proportion of the total variance of the test score (or indicator) that is attributable to the construct. For example, if a theorist defined the trait of creativity as independent of another trait (say intelligence) and the two traits correlate +0.40, then at least 16% of the reliable test variance is irrelevant to creativity as defined. A finding of this kind forces a redefinition of the traits, or a reconsideration of the methods used to operationalize them in the settings employed.

The MTMM mode of analysis developed by Campbell and Fiske (1959) emphasizes the point that to establish the scientific validity of a trait it is necessary to use some method to measure it. Each such measurement involves a trait-method unit, with the proportion of the true score attributable to the trait considered to be the true variance, and the proportion of the score attributable to the method to be systematic error variance. Further, it is necessary to measure more than one trait, and to use more than one method in order to determine the discriminant validity, by demonstrating that each of the different methods used can differentiate between independent traits. Convergent validity must be established by demonstrating that the different methods all measure a given trait in a similar manner. Unless such niceties of measurement are observed, it is difficult to elaborate a theoretical network because the core terms of the theory (T) are constantly beclouded by influences due to method factors, such as the auxiliary hypotheses underlying the measuring procedures and devices (A), and the specific conditions of the experiment (C).

For example, perhaps one is interested in the relationship between race and IQ, and postulates there is no systematic relationship. The outcome of the research program cannot be brought to bear on the core theory without some ambiguity. Assume that the IQ of a sample of urban blacks and whites is tested with the Stanford-Binet IQ test. If the results indicate that there is a significant difference between the IQ scores of whites and blacks it is doubtful that the theorist will reject the core theory because of the observed difference, nor should it be rejected. Many believe that the assumptions underlying the 1Q test might render it inappropriate for use in this context, and that a different, culture-free, IQ measure might be more appropriate, or that the underlying differences between the samples used and the standardization population might be crucial, or that the concept of race as embodied in the study is biologically suspect, and so on. In short, there are a host of methodological and conceptual problems that make it impossible to decide whether the negative results cast doubt on the core theory, are produced by inadequate measurement operations, or cast doubt on some of the auxiliary hypotheses underlying the choice of methods. Also, it is not possible to determine probable effect sizes of different variables because most reliable variables have shared variance; if this correlated variance is not partialled out, a misleadingly large estimate of effect sizes results.

Hammond, Hamm, & Grassia (1986) extended the MTMM procedures to study both the ecological and functional validities in human judgements. They also considered the results of two major research programs in cognitive psychology, and demonstrated severe problems concerning the adequacy of the theoretical generalizations that can be made from these programs due to the incompleteness of the research designs.

Hammond, et al. used MTMM to study characteristics of highway design. The method used is unusual, the problem unfamiliar, and the conclusions so striking that the study will be presented in some detail. The questions involved design aspects that would affect three “traits” of highways: safety, capacity, and aesthetic quality. The investigators chose 20 highway engineers and had them judge the safety of 40 highways. Each engineer judged highway safety (using a rating scale), capacity in terms of cars per hour, and aesthetic quality (again, using a rating scale). Each engineer made these judgements using each of three methods: (1) Intuition (based on viewing filmstrips of 1-3 mile segments); (2) Quasirationally (from bar graphs of nine attributes for each of the three qualities); and (3) Analytically (they constructed a mathematical formula for each concept). The data for the 20 engineers were combined to provide a result that would apply to an aggregated “artificial engineer.”

In addition, there were empirical criteria to evaluate the functional adequacy of the engineers’ judgements. For safety, the average accident rate over seven years was used; for capacity, calculations from the Highway Capacity Manual were used; and for aesthetics, 91 citizens judged the attractiveness of highway segments by rating the filmstrips.

Test-retest reliabilities were determined for each method and trait combination and were found to be acceptably high. Without going into further detail, it was found that there was high discriminant validity (the correlations between measures of the different traits by the same method were quite low), and there was high convergent validity (the correlations between measures of the same trait by different methods were considerably higher). These results indicate that the methods were adequate to capture the policies the members of the panel used to make their judgements. This initial analysis established the coherence of the measurement scheme, which is crucial to demonstrate construct validity on the grounds of internal logic.

After the above MTMM analysis was completed a similar analysis was performed adding the empirical criterion variables. The criterion values inquire beyond the level of internal consistency and probe the validity of the judgements of the panel members. The validity criteria were chosen to provide agreed-upon indices of each of the highway characteristics. By correlating the judged characteristics with "true" values, as indicated by the criterion scores, the validity of the judgements can be established–the functional validities can be estimated.

The value of the approach advocated by Hammond, et al. is not limited to the solution of practical problems. Proper measurement procedures should be at least as important in the development of scientific theory as in the solution of practical problems. Hammond et al. (1986) examined two major research programs to evaluate the adequacy of the data used to support the generalizations made. The research programs of Anderson (1974) Posner (1969) are both directed toward understanding memory for abstract and concrete knowledge. Anderson (1985) discusses the experimental evidence from these two sets of studies, and concludes that we remember abstract information and not physical details. However, an analysis of the experiments indicates that this conclusion is not warranted because both series of studies evaluated performance using a single response measure, reaction time in relation to interstimulus interval. Reaction time is a time-honored measure employed in memory studies. However, there are demand characteristics imposed by the reaction time method that would not be imposed by other methods, such as locating stimuli in different quadrants of a visual display, or in free recall. The point is not that reaction time is a poor measure, only that it could have unique characteristics that influence systematically the outcome of the research program. At the very least the magnitude of such influences should be determined relative to those of other methods.

Posner used a perception task and Anderson a verbal one, but neither established either the convergent validity or the discriminant validity of their procedures, nor did the subjects appear in more than one experimental condition, a fact which makes it impossible to assess reliabilities and difficult to interpret validities. Due to the incompleteness of the experimental designs, lack of reliability estimates, no attention to discriminant or convergent validities, a failure to isolate method variance, as well as a failure by Kolers (1979) to obtain the same pattern of results when using a different performance measure, no clear generalization can be made on the basis of the data available.

To be sure, it is difficult and time-consuming to conduct a study using a complete MTMM design. However, there is little hope that a progressive research enterprise can be based on confounded, piecemeal studies. At the very least we should take Campbell's (1986) suggestion seriously to the effect that the central findings of research programs should be replicated using different methods insofar as theory does not specify method. The hopelessness that some behavioral scientists have accepted as a given is expressed in the following comment received by Ken Hammond (pers. comm.) from a prominent psychologist, "I think you have a fine paper here. I still doubt that experimentalists like myself will be influenced by it, but that reflects our unwillingness to do studies of the magnitude required (and, I fear, publication pressures) more than the cogency or clarity of your arguments." It is sad to consider such a statement by a reputable scientists in the light of the immense magnitude of time and effort that has been expended conducting almost uninterpretable, but highly publishable, studies.

Meehl (1990, p. 218) summarizes the deplorable state of affairs that results from attitudes such as the above as follows: "It may be objected that it would be too onerous to require that investigators plug in a whole bunch of things that they ought to be worried about with the Campbell-Fiske discriminant validation in mind. All I can say to that is that, absent a tradition of so doing, I do not know how much confidence to have in detached validity claims for testing substantive theories." He suggests (1990, p. 198) that this intolerable state of affairs continues because "...students and colleagues have trouble hearing [these claims], since they might not know what to do next if they took it too seriously!"

It should be made clear that, although individuals might be unwilling to conduct research programs in a manner that will support intended generalizations, the collective community of scientists, each doing some of what needs to be done, will produce an acceptable outcome. A similar view can be applied to the falsification process. Research has indicated that individuals do not use negative evidence nor seek disconfirmatory evidence when solving logic problems in the laboratory, nor do they attempt to evaluate evidence in terms of alternative hypotheses (Wason, 1968). Among scientists there clearly is a tendency to avoid giving up a favored hypothesis. Yet, the scientific community, as represented by general theorists and authors of review articles, does employ severe falsification procedures, and evaluate existing evidence in terms of alternative hypotheses. Thus, the human propensities that lead us to nurture and cherish the scientific ideas to which we have given birth do not lead to an inevitable lack of scientific progress through the continued survival of unfit conceptual models. As Hull (1988, p. 343) phrased it, "Falsifiability...concerns theories, not theorists; science, not scientists."

Ecological and Functional Validity

Another pervasive problem occurs because of a failure to consider the distinctions Brunswik made between ecological and functional validity. The distinctions have been developed in detail elsewhere (Petrinovich, 1979, 1981, 1989), and will not be discussed at length here. In Brunswik's (1952) terms, ecological validity refers to the degree of trustworthiness of proximal stimulus elements to mediate (or signify) distal events.

Thus, in the stimulus domain, ecological validity refers to the structure of the environment, and such validity can only be known through careful and painstaking analysis of the relationships existing between distal and proximal stimuli. Functional validity refers to the organism's *use* of the structure of environmental stimuli. An organism might not use a stimulus (it is assigned low functional validity) even though it is a good stimulus (has high ecological validity). Stimuli also differ in terms of their reliability, both ecological and functional.

There is an almost universal misunderstanding of Brunswik's concept of ecological validity; it does *not* refer to the naturalness of a research setting, but is a technical, specialized term within the Brunswikian system. As indicated above "ecological validity" refers to the potential *utility* of various cues for organisms in their ecology, while "representative design" refers to the quality of naturalness, or lifelikeness, of the research.

Hammond (1978) pointed out that the Brunswikian meaning of "ecological validity" was established for three decades (1947-77), but that its meaning has been eroded. People now speak in terms of the ecological validity of an experiment when they really mean the representativeness of the design (e.g. Banaji & Crowder, 1989; Bandura, 1978; Bronfenbrenner, 1977; Gardiner, 1990; Neisser, 1976; Neisser & Winograd, 1988; Parke, 1976). Misconstruing the technical meaning of ecological validity, and its distinction from functional validity, can lead to problems that are based on a failure to appreciate the difference between objective, material features of the environment and the organism's subjective construal of those features (for an example of such a misconstrual see Stokols 1982). As I indicated elsewhere (Petrinovich, 1989, p. 14), "Perhaps it is enough that people are now concerned with the problem of representativeness, and the technical Brunswikian term should be surrendered." However, the distinction between ecological validity and functional validity is a valuable one and should not be obscured because each of these validities focuses on a different and crucial aspect of the behavioral situation. Ecological validity inquires into the structure of the environment and functional validity into the utilization of that structure by the behaving organism and this is a crucial distinction to maintain.

External and Internal Validity

Campbell (1957) introduced a much discussed distinction between internal and external validity. He defined internal validity as a concern with the question, 'did the experimental stimulus make some significant difference in this specific instance?' External validity he equated with representativeness, or generalizability and posed the question, 'to what other populations, settings, measures, treatments, and times can an obtained effect be generalized?' His concern was with the extent to which the controls required for internal validity tend to jeopardize the representativeness of the study, and thereby jeopardize its generalizability. Campbell (1986) has subsequently argued that we should make a clear distinction between the validity of theoretical interpretation and atheoretical generalization to other samples from our intended universe of generalization.

Recently, there have been a number of challenges to the ideas of those who emphasize the importance of external validity, and the questions in general revolve around the relative merits of systematic and representative design. The issue is, can one build an adequate science of behavior through a reliance on the single variable ideal of the laboratory model of science, or is it necessary to include the logic of representative sampling? The issues revolve around the question of whether or not it is necessary to be concerned with questions regarding external validity (representativeness), and whether results from the laboratory are sufficient to support a theoretically sound science.

One frequently cited paper is that of Mook (1983) who argues the merits of the artificiality of laboratory settings on the grounds that generalizations to the real world may not necessarily be intended. He maintains that findings in the laboratory have added force because of the artificiality of the setting: "...we may demonstrate the power of a phenomenon by showing that it happens even under unnatural conditions that ought to preclude it." (p. 382) Mook (1989) and I (Petrinovich, 1989) have debated these points in print and I will only touch on a few issues that still trouble me because the views expressed by Mook continue to be endorsed in the literature.

Mook (1983) discusses Harlow's research on "mother love," and Banaji & Crowder (1989) characterize this discussion as an emphatic argument in defense of external invalidity. The essential argument is that this work is regarded as a major contribution, yet the monkeys were not representative of any natural population of monkeys, nor was the laboratory setting representative of any ecological setting. Mook (p. 381) asks, "What can this contrived situation possibly tell us about how monkeys with natural upbringing would behave in a natural setting?...to make generalizations about real-world behavior was no part of Harlow's intention....What

Harlow did conclude was that the hunger-reduction interpretation of mother love would not work. If anything about his experiment has external validity, it is this theoretical point, not the findings themselves....We do not dismiss the findings and go back to do the experiment "properly," in the jungle with a random sample of baby monkeys." (p. 381)

Mook concludes that "...it is [theories] that generalize to the real world if anything does." (p. 383) Here is the crux of the problem: A theory is a general statement about some specific set of occurrences, and always applies to some universe of occurrences. It can be conceived to apply universally to all organisms, to just humans, to hold widely, to apply only to certain individuals at certain times and places, or to apply only to this individual in this place. The question of sampling representativeness always rears its ugly head whenever a theory is about *anything*, and clearly all theories are about *something*. If so, then that something must be very clearly specified.

Mook (1983) considers a study by R. Brown & Hanlon (1970) to demonstrate that there is no necessity to representatively sample subjects. I discussed this issue elsewhere (Petrinovich, 1989) and concluded that, because the theory of concern was Chomsky's theory of universal grammar, which argues that language involves universal processes, there was indeed no necessity to sample subjects because the theory is cast in such a way that it holds for all normal humans. There was a necessity to sample sentence types, and to be concerned with the representativeness of methods of recording parent-child interactions, however, and Brown and Hanlon did show great concern to sample those aspects that would be variable in terms of the theory under consideration. Because the theory applies universally, the only question one need answer concerns the reliability of the observations of syntax development.

In 1989, Mook and I exchanged views, and he wrote "This [external validity] we can define for our purposes simply as *the extent to which experimental findings make us better able to predict real-world behavior*." (p. 25), which is his construal and not the position I am advocating. I have taken great pains to argue that generalization to anything is acceptable, as long as the information at hand bears on the populations relevant to the theories under consideration. "The issue should be considered from the perspective of adopting sampling strategies that will support the scientific generalization to be drawn: Do the sampling procedures support the inference from the particular samples chosen to the universes of interest?" (Petrinovich, 1989, p. 11).

I would have been content to let the matter drop had not yet another article appeared (Banaji & Crowder, 1989) which pursues the same vein as that mined by Mook. Again, the crux of the article is to argue that concerns regarding ecological validity (and, here, they mean representativeness) are best ignored. They argue that the scientific method, as they narrowly construe it, is the proper method to arrive at "...the empirical discovery of facts about memory that have *generalizability*, and not by the use of tasks that carry an illusion of ecological validity by testing memory in everyday contexts." (p. 1187)

They continue that (p. 1189) "...the multiplicity of uncontrolled factors in naturalistic contexts actually prohibits generalizability to other situations with different parameters." But, I am at a loss to understand how we identify whether or not there are different parameters or how we determine that they are important unless we study the parameters in the universes to which we wish to generalize. The preceding argument of Banaji and Crowder, taken to its logical absurdity, would suggest that we can never generalize to anything.

Again, a straw man is created by constructing a "Two-by-Two Array of Approaches to Science," with the two dimensions being "Ecological Validity of Method" and "Generalizability of results." They argue that "ecologically validity oriented" scientists would opt for the High Ecological Validity of Method cell, "lifelike methods at all costs." I know of no such scientists myself: there are those of us who want to inquire into the reasonableness of, and problems involved in, extending findings beyond those conditions under which they were obtained, but that is a far cry from "lifelike methods [or real life] at all costs."

The rhetoric used in this article is sprinkled liberally with hyperbole and ridicule and contributes little to the solution of the serious problems relating to the adequacy of theoretical generalizations based on evidence. I agree when they write, "We do not wish to condone smugness about the generality of laboratory principles to any external context. In fact, we need to test these applications assiduously." (p. 1191) The importance of testing generality of laboratory principles is what I am advocating here, and I would add that that is exactly the position critics such as McGuire (1973), Hammond (1978), and Funder (1987) have been advocating within the realm of social psychology. Banaji and Crowder characterize such critics as "alarmists" who speak of the crisis in social psychology and argue that "Social psychology... must be concerned with real events and real people if it is to comment on the nature of social behavior." (p. 1192) Rather, I believe these social psychologists are methodologists arguing that generality must be established scientifically, and that they have sought to test these application studies "assiduously," exactly as Banaji and Crowder have recommended.

Null Hypothesis Testing

Arguments regarding the inadequacy of null hypothesis testing procedures have been accepted by statisticians and methodologists for over 40 years, and it should hardly be necessary to argue the case yet again. However, the use of null hypothesis testing is still the dominant practice among researchers in the behavioral sciences, we still teach it to all of our students from introductory statistics courses onward, and some statement of significance level is insisted on by almost all journal editors. Yet, statistical significance testing not only is inadequate to provide support for a progressive science, it is harmful.

Various writers have developed the critical arguments, and many of the early papers have been collected in an edited volume by Morrison & Henkel (1970), the points have been summarized by Meehl (1986), reviewed and updated by Oakes (1986), and placed in a more philosophically oriented context by Serlin & Lapsley (1985). Oakes (1986, p. 43) states the case succinctly, "If after 40 years of significance testing we are not ready to identify our theories with the null hypothesis, what grounds exist for supposing that a further 40 years' significance testing will do the trick? Of course, precise predictions are only a distant prospect precisely because we have not been building up repositories of reliable estimates of effect sizes, but have been speciously accepting and rejecting theories that are barely worthy of the term."

I will summarize a few of the salient points in the argument, assuming a familiarity with the steps involved in testing null hypotheses and establishing significance levels.

(1) There has been a tendency to confuse the decision regarding statistical significance with substantive confidence (degree of belief). Rejection of the null hypothesis is a statistical decision: the classic Neyman-Pearson model involves the assumption of a value for the population parameter (null difference between means) and estimation of the probability that the obtained sample value falls within acceptable limits of that population parameter. Confidence, however, is a cognitive process, and relates to the strength of belief in a scientific hypothesis. Arguably, the best way to increase confidence is to elaborate scientific theoretical networks by developing multiple alternative hypotheses, using sound measurement models, and attending to issues affecting the quality of generalization. We should then proceed through the further development of alternatives to those theoretical networks, rather than making rely on making "yes" or "no" decisions. As Meehl (1986, p. 324) wrote, "The important thing to clarify is the structure of the theoretical network and the resulting empirical tests." and Harcum (1990, p. 405) adds, "The problem should be solved by more attention in research to overall relationships and replications."–a point I will emphasize later.

(2) The classic decision rule procedures constrain us to establish a test value, and if the test statistic attains that value, to accept or reject the null hypothesis with "the blessing of automaticity," as Bakan (1967) so nicely phrased it. By doing this, one is in the strange position of being

forced to decide that a t statistic of 1.96 and one of 1.95 are totally different: In the first case we might have to accept the null hypothesis, and in the second we might have to reject it. It can be argued that no scientists really behave this way, nor should they. Rozeboom (1960) has argued that we really do not take it all seriously, anyway. If a p level is 0.04 as opposed to 0.06 we proceed according to our beliefs regarding our substantive hypothesis, publish, and continue research. If we find a value of 0.06, and still believe in our substantive hypothesis, we add more cases or "improve" the study, and run it again; if we do not believe the hypothesis we claim support for null at 0.06, and develop a substantive argument to account for the failure to reject.

When one considers the multivariate analyses that are now available, the question of significant increments in variance accounted for often becomes trivial. When using stepwise multiple regression procedures, or causal modelling techniques, one can obtain statistically significant increments in the size of an R^2 (which indicates that a significantly greater proportion of variance is accounted for) or a reduction in χ^2 (which indicates a better fit between the model and the data) when more and more estimates are added. However, a rule based on the rate and size of the increments in effect size proves to be more reasonable, valuable, and practical: minuscule increases in effect size, even though they have high levels of statistical significance, unduly complicate understanding of the causal texture of the variables at play, and with too many terms we run the risk of doing little more than describing the obtained results. As the old saw goes, given a large number of constants we can fit an equation to a dog and, with a few more, make the tail wag.

(3) The importance of developing alternative hypotheses has been discussed in broad terms, but the value of such development is especially valuable when null testing procedures are used. As indicated above, rejection of null is often interpreted as strong corroboration of a substantive hypothesis, whereas it really is quite feeble, because the likelihood of the alternative, substantive hypothesis, is never determined; only the "unlikelihood" of the null. As Harcum (1990, p. 404) reminds us "...a null result can be virtually guaranteed by imprecise research...." Cochran & Cox (1957, p. 5) wrote, "In many experiments it seems obvious that different treatments must have produced some difference, however small, in effect. Thus the hypothesis that there is *no* difference is unrealistic: the real problem is to obtain estimates of the sizes of the differences."

Accompanying the extreme reliance on null hypothesis testing is a misinterpretation of the meaning of the different probability values. There is no direct linear relationship between effect size and significance level. A result that is significant at the 0.01 level might be thought of as five times as significant as one at 0.05, in terms of the likelihood of observing it if null is true. However, the increased level of significance here does not translate into five times the explained variance. Oakes (1986) pointed out that this factor of five is not found among measures of effect size. If an appropriate measure of effect size is calculated, the result at 0.01 accounts for only about 1.7 times as much variance as one at 0.05, not five times as much. It must be emphasized that test statistics are a function of both effect size and of sample size. A statistically significant test statistic may be due to the detection of a large effect with a small sample or a small effect with a large sample. Funder & Ozer (1983) have demonstrated convincingly that there is no simple direct relation between size and importance of effects and that considerations regarding the theoretical and practical relevance of effects, should exercise the directive role in evaluating the importance of effects.

Replacing requirements that significance levels are met by setting a required minimum value for effect sizes as the criterion by which research will be evaluated is not an adequate solution. This replacement would just involve a reliance on one type of automaticity in place of another. Effect size can be manipulated through a biased selection and control of variables, by a judicious choice of scalar values for manipulated variables, improper sampling of settings, or an arbitrary selection of behaviors chosen for study. To obtain ecologically meaningful estimates of effect size, the principles of representative design (Brunswik, 1956; Petrinovich, 1979) must be invoked. When this is done, theories based on sets of observations can be generalized to the universes of generalization appropriate to the behavioral ecology of the organisms in question, and the effect size estimates reflect the probable importance of the variables to account for variance in behavior, and not simply their possible importance.

In spite of these obvious points, there still exists a tendency to interpret statistical significance as a continuum. For example, Kanekar (1990, p. 296), believes that, “A result of higher significance is a result that inspires greater confidence in the rejection of the null hypothesis.” and this is the problem: rejection of the null refers to that hypothesis only. With no consideration of effect size, one cannot determine the importance of the variables embedded in the alternative hypothesis. That it does so is unfortunate given the fact that so many factors, other than the size of the effect, can influence the size of a significance level.

Chow (1988), in an article in the Quantitative Methods section of the leading APA journal devoted to such matters, arrived at the conclusion that nothing is gained by using an effect size estimate rather than a binary decision of whether one group is different from another. This surprising conclusion is based on a faulty syllogism that does not allow for the fact that, when an experimental outcome is not as predicted (say, two groups are not significantly different from one another), it is impossible to know what aspect of the theoretical fabric is incorrect: is it (T), the core theory; problems in the measuring instruments and their auxiliary hypotheses (A); or the specific conditions of the design (C)? Chow argues it is proper and useful to use null hypothesis models when "Theory-Corroboration Experimentation" is being conducted. He employs the *Modus tollens* to evaluate a theory (T) (Table 2, p. 107): First, he states a major premise–given the assumptions of the experiment (A), then we have an expectation (X), given the types of control and independent variables used in the experiment (EFG); He then assumes a minor premise–the experimental outcome (D) is dissimilar to that expected (X); This leads him to an experimental conclusion–that the assumption underlying the experiment (A) is false; Finally he draws a theoretical conclusion–the basic theory (T) is false. (Or, in the terminology used in this paper, T leads to a prediction (P) given the conditions A.C.)

The problem with his argument is that the major premise is false: if the predicted D does not occur he concludes that the core theory is false. However, the major premise is not a single unit: we do not know if T is false, A is false, C is false, or some combination of T.A.C is false. The substantive theory could still be true, but one of our auxiliary hypotheses (of which there are always a large number) could be false, or the specific experimental conditions chosen are not adequate to the task of measuring the dependent variables relevant to T. Thus, we know not what we should accept as false, and little progress has been made toward theoretical understanding or explanation. Chow's views seem limited to a narrow view of theory, and, if applicable at all, apply to single-variable experiments that are concerned with justification, primarily. As I have argued, the context of discovery is often of far greater importance if we are to realize scientific progress, and there is little hope of learning much following Chow's strictures. Chow (1988) expresses disapproval of the statement, "Although the effect is statistically significant, it is nonetheless very small." He considers the statement misleading because it misrepresents the binary nature of the appropriate statistical decision, and argues that the theoretical expectation should be a qualitative one (Is D like X?), and not a quantitative one (How unlike D is X?) He finally states that, "...all that is required of a statistical analysis is a binary decision. This is the case because the validity of the syllogistic argument requires only that information. Even if a quantitatively more informative index is available (e.g., effect size, the amount of variance accounted for, or the power of the test), it will still be used in a binary manner. That is, nothing is gained by using an effect-size estimate in this context." (p. 108)

As indicated, the syllogism is not valid; the T.A.C conjoint defeats the major premise. We do not have a binary decision, and if we do make a binary decision, why should we not use the additional information that is embedded in the experimental data? Although, properly, he indicates that effect size can be influenced by the choice of experimental manipulations that may be removed from the theoretical property of the underlying mechanism, the rigorous implementation of representative design will circumvent many of the problems that occur as the result of measurement issues.

Chow (1988) concludes his argument (correctly, in my view) that a theory is strengthened, not by mere literal replications of the same experiment, but by a series of converging operations. The context of discovery must be served, in Chow's schema, by a series of converging operations that depend on a series of *constructive replications*. While performance of such replications is a commendable aim, the overall view offers a slim hope of developing a progressive research program given the continued reliance on null tests.

(4) Upon careful reflection, it becomes obvious that the null hypothesis is always false. The likelihood of rejecting the null is influenced greatly by arbitrary choices of level of significance, arbitrary design decisions, which have been discussed above, the number of subjects chosen for study, and the true size of the population deviation from null.

Meehl (1967) has presented a compelling thought experiment which makes the point that the true probability of rejecting the null at the 0.05 level, with a non-directional test, is really closer to 0.5 in the true null case. Meehl's arguments support the conclusion that the main reason we might not reject the null is a lack of power in our tests, either due to an insufficient number of subjects or to unreliability of measurements. The implication is that as we do more adequate studies in terms of obtaining large random samples, and improve the validity and reliability of our measurements, our theoretical models become subject to less and less strict evaluation: the null hypothesis barrier becomes easier and easier to surmount. If any of the variables we are manipulating control even a minuscule proportion of the variance in the population, we will be able to reject the null hypothesis with greater and greater ease. Thus, as measurement becomes more precise and reliable, and control becomes better, the empirical criterion for theory to satisfy becomes increasingly weaker. This circumstance, alone, can make it almost impossible to develop strong theoretical networks.

Meehl (1967) considers this a paradox: In the physical sciences, with improvement in experimental design, instrumentation, or mass of data, there is an increase in the difficulty of the "observation hurdle" for theory. With such increases in precision, it becomes possible to make more accurate and particular predictions, these riskier predictions provide a stronger corroboration of theory if they are supported, and it becomes more apparent at what point predicted functions do not obtain. In behavioral science, such improvements provide an easier empirical hurdle, providing a weaker corroboration of theory, and less strict evaluation of theoretical principles because we use the null difference as the hypothesis in our tests of significance.

In addition to logical arguments regarding the weakness of null hypothesis procedures, there have been empirical demonstrations that the null hypothesis will almost always be rejected when the number of subjects tested is large (e.g., Bakan, 1967).

Meehl (1967) refers to a study by Lykken and Meehl in which they examined the intercorrelations of data based on 45 miscellaneous variables gathered from 55,000 Minnesota high school seniors that were examined. Ninety-one per cent of the pairwise correlations between any pair of variables were significant, and the majority of variables exhibited significant relationships with all but three of the others (often with $p < 10^{-6}$). The 9% of the correlations that were not significant were based on measures of dubious reliability.

Nunnally (1960, p. 643) wrote, "If the null hypothesis is not rejected, it is usually because the N is too small. If enough data are gathered, the hypothesis will generally be rejected. If rejection of the null hypothesis were the real intention in psychological experiments, there usually would be no need to gather data." More recently, Murphy (1990, p. 403) has echoed this sentiment, "If one assumes that the null hypothesis is never true, the practice of testing the null is absolutely uninformative; its results are a foregone conclusion....the set of alternatives to be tested (i.e., H_0 vs. H_1 should contain hypotheses that have some non-trivial probability of being true."

(5) One seldom discussed point is that unless we have a rule concerning when we should stop adding subjects to an experiment there is no doubt that null will always be rejected. Assume we have two investigators. One gathers pre- and post-test data using 20 subjects, chooses to calculate the t statistic to evaluate the data, and sets the significance level at 0.05. This investigator finds a t ratio with $p > 0.05$. The decision should be, then, to abandon the problem because the treatment did not produce a value in the region of rejection. However, because the t ratio was "very close to significance," the investigator decides to add 10 more subjects. The results based on only these ten additional subjects fails to yield a t ratio for which $p < 0.05$. However, if the data obtained with these ten subjects, which were treated the same as the previous 20, are combined to give a total N of 30, then. the t ratio supports a rejection of the null at $p < 0.05$. This post-experimental analysis is not legitimate when using the Neyman-Pearson decision theory. If one continues to sample from a truly random universe until a significant result is found, and stops when it occurs, there is no doubt that null will always be rejected. One will be able to capitalize on the short-term runs that occur in the desired direction, and stop when that

run has taken place: a comparable run in the other direction, which would be expected in a truly random universe, is prevented from occurring. Further, let us consider a second investigator who decided to run all 30 subjects at one time, and who obtained identical data. It is legitimate, here, to consider all 30 subjects in one analysis, and the identical data as found by the first investigator legitimately will support the decision to reject the null at p <.0.05. A disturbing thought is, would it be legitimate to accept the results found by the second investigator if the results of the first 20 subjects merely had been glanced at without actually computing the means, variances, and t ratio, and, following this informal examination, the additional ten were tested? As Oakes (1986, p. 116) points out, it is "...ludicrous that the furtive behaviour and intentions of a scientist should influence the evidential import of his data."

Clearly, the interpretation of the meaning of experimental data should not rest exclusively on the decision to accept or reject the null hypothesis: close attention should be paid to such things as effect size when sample size is increased. With an increase in N, a much smaller effect size will result in a decision to reject null. However, in such cases the experiment becomes less and less interesting. For example, with 1000 degrees of freedom a Pearson correlation coefficient only has to be 0.06 to reach the 0.05 level of significance; yet a coefficient that size accounts for only 0.0036 of the variance in Y, and an effect that small is seldom of practical or theoretical interest.

Even this abridged account of the perils of relying on tests of the null hypothesis make it evident that the procedure is not sufficient to support a progressive research program. I endorse the view expressed by Oakes (1986, p. 66) that, "...the continued use of significance tests does nothing to encourage the development of rich causal theories capable of producing non-trivial statistical predictions."

The direction of science is determined primarily by human creative imagination, and not by the universe of facts which surrounds us. The facts we choose as relevant are selected arbitrarily, and their place in the theoretical fabric is often post hoc in nature. There is no effective falsification before the emergence of a new theory. These concerns lead to the conclusion that, because justification is logically of minor value, methods should be emphasized that enhance discovery–the development of risky theories and the embellishment of useful auxiliary hypotheses–and the proposed theories must be subjected to severe tests.

Methods should be developed that enhance the context of discovery–ones that employ pluralistic yet rigorous methodology. The need is for a logical, rational context of discovery: a methodological system that is adequate within the context of justification, without abandoning the possibility of discovery. The foregoing arguments indicate that reliance on a single hypothesis, referring to a single construct, that is measured by one method and evaluated using hull hypothesis tests, will not allow us to achieve the methodological ideals that have been discussed. Over and again it is said "We all know this," and yet the vast majority of behavioral scientists still adhere to these procedures.

Is There A Cure?

This section is based on the premise that there is a need to develop an alternative approach to the development of scientific theory. Some solutions in the realm of measurement will be suggested that might enable us to progress more rapidly in the attempt to develop more adequate theory. Meehl has come to much the same conclusions as those expressed here regarding the state of what he calls "soft psychology," and argues that we must find ways for weak theories to be tested strongly. He also considers it most important to clarify the structure of theoretical networks and to evaluate carefully the resulting empirical tests. Meehl (1986, p. 325) clearly summarizes the situation, "The distribution of obtained significant and nonsignificant results is an arbitrary and complex artifact of eight methodological factors largely unrelated to a theory's verisimilitude, namely, (a) experimental design, (b) inherent construct validity of measures, (c) reliability of measures, (d) properties of the statistical power functions, (e) presence and size of higher-order interactions, (f) verisimilitude of auxiliary theories relied on in deriving higher-order interactions, (g) differential submission rate of manuscripts reporting significant versus nonsignificant findings, and (h) editorial bias as to the same. The net result of these influences on the pro/con count is that usually such a heap of studies is well nigh uninterpretable."
Meehl states that he hesitates to paint such a bleak picture without having a clever and convincing "cure" up his sleeve, but regrets he is unable to provide one. I will, in the next section, attempt a sketch of what I hope is a step toward such a cure. It will be useful to identify the nature of the treatments that might be applied to control some of the symptoms before writing the research prescription intended to cure the illness. (1) I have argued above that many of our problems are due to a determined effort to avoid issues relating to adequate design and measurement aspects of our research. I have discussed the importance of establishing construct validity, and using the MTMM mode of thinking when undertaking a program of research. In addition, the problem of low reliability of measures is seldom faced in behavioral research, especially that which is done in laboratory settings. Epstein (1980) forcefully presented the case for aggregating behavior over situations and occasions to cancel out incidental and uncontrollable factors, and to achieve reliable estimates of behavioral entities.

Block (1977) found impressive coherence and consistency of behavior and personality in human development by utilizing methods to increase the reliability of measurement of traits at different stages of development. It is apparent that much of our lack of progress in developing progressive research programs is due, not so much to weak attempts at building theory or designing experiments, but to faulty and slip-shod decisions regarding issues involved in measurement of the observable behaviors that serve as indicators of constructs.

(2) The basic statistical model used should be changed. Seldom is a hypothesis of no difference between groups adequate to the task of developing theory. Yeaton & Sechrest (1987) discuss several historical events in terms of "no difference research" paradigms, ranging from the possibility of extraterrestrial intelligence, the existence of the Loch Ness Monster and Bigfoot, regarding all of which the null has been difficult to prove to everyone's satisfaction. They contrast this with Captain Cook's proof, obtained through circumnavigation, that there was no Southern continent. To make a long, and delightful, story short, the problem with proving the null is that the null has to be stated unequivocally in terms of finite space and time, and the operational indices that would be adequate to prove or disprove the hypothesis have to be precisely formulated. They argue that the null is provable, but only if you are willing to state with some exactness the conditions of its acceptance and to state the exact conditions under which we will give up our belief in a hypothesis.

Another simple improvement is to develop estimates of likelihood ranges, to establish confidence intervals, using the obtained sample value as the estimate of the population value, and to establish the likelihood range of the "true" value based on the obtained sample. This is preferable to the method of assuming a nil population value and establishing a range around that value. As Oakes (1986, p. 52) points out, "The significance test relates to what the population parameter is *not*; the confidence interval gives a plausible range for what the parameter *is*." It is also clear that one can have one's cake and eat it too when comparing results using, say, a distribution of mean differences: if the estimated range of the population difference does not include zero, the null hypothesis, thereby, has been rejected; the confidence interval, used in this way, gives all of the information the traditional significance test does, plus a great deal more.

(3) A strong suspicion that liberties are taken when testing the null hypothesis receives support from the fact that the null is rejected over and again when the power of the tests used is *a priori* very small. Cohen (1962) examined all articles published in the *Journal of Abnormal and Social Psychology* for one year and found that 70 studies reported significant results. When he calculated the power of the tests used, the power to reject null was quite low. Yet, in all cases the null was rejected. He suspects that

the results of all experiments available were not represented in this sample.

Sedlmeier & Gigerenzer (1989) inquired into the possible impact that Cohen's analysis had on the power of studies published 24 years later in the same journal. In only two cases were any remarks made concerning power by the authors of the experimental studies. No author discussed why a certain alpha or number of subjects was chosen, or what effect size was expected. When the median power of the tests reported was calculated it was found that power had not increased beyond what Cohen found to be the case 24 years earlier.

They also report that in seven experiments the null hypothesis was stated as the research hypothesis. None of these tests were significant, and this result was unanimously interpreted by the authors as a confirmation of their research hypothesis. The median power of these nonsignificant tests was only .25. This means that the experimental conditions, such as number of observations, were set up in such a way that given a true medium effect, the research (null) hypothesis would nevertheless be "confirmed" in 75% of the cases (Sedlmeier & Gigerenzer, 1989, p. 313. They conclude, "This situation will not change until the first editor of a major journal writes into his or her editorial policy statement that authors should estimate the power of their tests if they perform significance testing, and in particular if H_0 is the research hypothesis." (p. 315)

Sterling (1959) surveyed the articles published in four APA journals in one year and Greenwald (1975) those published in one year in the *Journal of Personality and Social Psychology*. Both found results that support the conclusions reached by Cohen and by Sedlmeier and Gigerenzer. These finding all lend credence to the suspicion that null results are not often published, either because editors tend to reject papers with null results, or papers with null results are not submitted. Thus, the probability of Type I errors might well be much larger than stated. Forbes (1990) has pointed out that the same problems occur in the field of ornithology, and discusses instances in which a failure to reject null is used as confirmation of a substantive hypotheses, without any consideration of the question of the power of the test. Even more distressing are instances in which a lack of a significant difference is used to infer that a correlation is present between the two variables under consideration.

Rosenthal (1979) called the failure to publish null results the "file drawer problem," and computed an estimate of the number of null papers that must be filed away such that, when combined with the published results, the null hypothesis would just be rejectable at the 0.05 level. Although Rosenthal came to the conclusion, on the basis of a literature review, that the number that must be filed away was improbably large, Oakes (1986) convincingly challenged the basic premise on which the Rosenthal treatment was based, and arrived at a much more probable, far lower, number of “filed away” studies that would have to exist. All of the above findings suggest that Rozeboom was correct: we do not take the formal logic involved in the application of our inferential methods very seriously anyway, and I believe our science suffers incredible damage as result of this lack of concern.

Prescriptions for Design and Testing Theories

If, as usual, we have a vague substantive theory, we should employ the cures suggested in the preceding section, work through the logic and implications of our theoretical beliefs, and include every variable that can be observed in an initial study, as long as their inclusion does not damage the main thrust in the evaluation of the research hypothesis and the alternative ones which have been cast. We should, at the outset, critically review our conjectures and devise reasonable alternatives to our initial beliefs.

Relatively small samples of subjects should be used at the outset: the number of subjects chosen should be selected in the belief that they will provide reliable estimates of as yet unknown parameter values. This selection of a relatively few subjects accompanied with a large number of variables that we can submit to something like a multiple regression analysis violates one of the assumptions underlying multiple regression: there probably will be a disproportionate number of variables to subjects, but this is not serious in preliminary exploratory studies because constructive replications will continue, and these replications provide the required independent cross-validation of findings. These constructive replications will involve new operational translation of variables and sample other points from the universe of subjects, tasks, and situations to which the theory refers. There would be a serious problem only if the analysis was the final step in theory construction and evaluation. Campbell (1986, p. 76) advocates a similar approach when discussing methods for program evaluation, “...exploratory contrasts should be sought out for cross-validation that differ as much as possible from the first intervention in population, setting, and so forth while remaining within the...targeted populations and problems.”

Studies should be done that have a low, but reasonable, degree of power, perhaps a figure of one-half of a standard deviation in mean differences would be appropriate. At the outset we might have to guess at the expected effect sizes, because it is not possible to estimate power, lacking an estimate of the size of the standard deviation for the population. If there is a powerful effect and if we have reliably and carefully constructed measures of our variables, we should detect differences, say between two groups, by establishing confidence intervals for the range of differences. This small sample procedure should minimize the likelihood of detecting irrelevant non-null hypotheses that are based on weak effects because the limited (but reasonable) power of our test will not reveal the effect of variables that control small amounts of variance.

If a significant effect if found, another small sample experiment should be run that is a constructive replication of the first. In the interest of establishing a network of constructs, this replication should include some variation in the subject pool, the precise nature of the task, stimulus material, and response measures. If we choose these variations in a manner to assure that they are unbiased samples from their respective universes of generalization, such differences should not have a substantial impact on results.

Based on the initial study, we should use Cohen's (1962) strong effect of a one standard deviation difference, for example, between means, use a *p* level of 0.05, a conservative two-tailed test, and calculate the sample size required on that basis. If the new study, again, results in significant differences, we should keep varying conditions and including new variables until we do not obtain significant effects. When no effect is obtained it might be reasonable to perform literal replication of the study, as well as to guess at the difference making the difference, and include a probe for that difference in the new study.

This research gambit should permit the addition of new particulars with the greatest economy of subjects, time, and energy. Each constructive replication helps to overcome the problem of capitalizing on chance relationships, because each replication that gives the same pattern of results serves as a cross-validation of hypotheses on an independent sample. This process utilizes the comparative method to maximum advantage: At every step, we are including and excluding variables and comparing outcomes in the light of these changes.

What happens if there are no significant effects at the outset, but the investigator still believes there is something worth investigating? One could perform a literal replication to increase power before abandoning the hypothesis, and, for this replication, there will be preliminary data to estimate the necessary N to achieve the desired degree of power.

The constructive replications should be designed to elaborate the theory. These should be scaled in ecologically meaningful units, to estimate their effect size. It should also be possible to issue a set of *ceteris paribus* clauses: clauses that insist certain specified contaminating or perturbing influences are not present. Each such clause places a restriction on the extent of generalization and can, if the reasons for the perturbation are investigated, provide an additional research hypothesis to be investigated in order to understand why a given limitation exists on some parameter. The choice of variables, measures, and manipulations are not done atheoretically only to obtain representative samples of methods, subjects, and situations. The choice of terms for replication is done because of theoretical relevance. As Campbell (1986, p. 76) writes, “Purposive sampling for maximum exploration of generalizability on conceptualized dimensions will be substituted for population representative sampling...clinical experience, prior experimental results, and formal theory are very appropriate guides for efforts to make the exploration of the bounds of generalizability more systematic.”

Following this series of constructive replications and the refinement of the variable set, a proper multivariate design can be used, incorporating all of the important particulars discovered and moving toward better estimates of effect sizes and the nature of interactions between variables. One can also be able to partition the variance, to determine how much of the variance in the behavior included in our universe of generalization is explained and how much remains unexplained.

The ultimate goal is to generate point predictions concerning critical aspects of the developing theory. At this point the sampling strategy should change: large samples should be chosen which will result in a narrow confidence band around any predicted parameter value. This “no difference” strategy will result in a desirable high level of power, with the aim being to fail to reject the hypothesized parameter value based on the sample statistic. This recommendation is similar to the spirit of the suggestions by Serlin and Lapsley (1985) who argue for the use of a “good-enough principle” which essentially involves constructing confidence intervals around predicted values (not necessarily null, as they seem to recommend), using tests of high power, and setting strict limits before accepting the confirmation of the prediction.

The final step will be to fit curves, determine the nature and strength of interactions, and develop causal models using some procedure such as structural equation modeling (e.g. Jöreskog & Sörbom, 1981).

At this juncture, significance tests will be of little interest because we will have moved beyond the question of whether or not observed patterns of data are merely the result of chance sampling differences. Rather, methods are used that enhance the discovery process, and the aim is to build better theories, not merely to test existing ones. In this scheme, ideas lead, it is acknowledged at the outset that “facts” are arbitrary (being a set that are selected from the array of possible observations), and facts are meaningful only when they can be placed in the context of an explanatory model.

Although many of the above suggestions seem overly ambitious we should, as Meehl (1990, p. 233) suggests, “...be more optimistic about the possibility of making predictions beyond mere non-null difference predictions from rather weak theories.” Even if it is not possible to derive precise numerical expectations, it might still be possible to predict rough function forms: We do not have to wait until strong, solidly based theory is available to take the first steps toward developing experimental designs that will permit valid generalizations regarding the universes of events in which we are interested. If it is not practicable to develop full MTMM designs that will support generalizations based on data obtained in truly representative settings, at least one can utilize more than one subject pool, task, method, and situation, with stimuli and responses selected for study on the basis of their social and biological importance. Such partial steps might at least get us started on the way toward understanding the principles of behavioral science. As Meehl (1990) has outlined, positive, advances can be forced, not only by investigators taking to heart some of the points developed here, but also by journal editors and referees. Those individuals controlling access to the channels of scientific communication should insist that parameter estimates of obtained statistical values be provided (with confidence intervals), that reasonable estimates of the percentage of variance accounted for be included in research reports. Consideration of statistical power should be obligatory in the review of every negative result.

The entire enterprise outlined above is based on the argument that the currently used justification procedures are of questionable value, and that they stifle the discovery process. The research gambit suggested attempts to define and establish a universe of generalization and to explore the fabric of that universe of variables as fully as possible. Attention is given to significance testing (by way of examining confidence intervals) at the outset, but the significance testing procedures are not confused with substantive confidence. Attention is paid to the degree of power of statistical tests, with N's chosen on the basis of reasonable, but low, power at the outset, and care taken to obtain representative estimates of effect size.

A broad definition of variables is used that will tolerate a range of scalar values that have been chosen because they represent the range of biological, ecological, and social values found in those situations that characterize the universe of generalization, thereby making it likely that estimates of effect size will not suffer too greatly from distortions produced by unrepresentative range, density, and covariation of variables. The final step is to use powerful tests to subject our risky conjectures (risky because they contain precise predictions of function forms) to strict tests.

Epilogue

I suggested, at the beginning of this article, that the methodological and analytic strategies that are used in the behavioral sciences do not seem to be adequate to the task of supporting progressive scientific programs. I am endorsing a form of *instrumentalism* that considers theoretical terms to be mere instruments for organizing claims about the things referred to by evidence based on observational terms. I do believe, however, that we are not hopelessly lost in a circle of relativism, but that our theories increase in their verisimilitude because of our efforts to develop more adequate, inclusive, and heuristic theories.

Giere (1988) is quite correct when he emphasizes the fact that we cannot test a theory directly against observation, "Instead, the model must be embedded in an experimental context..." (Giere, 1988, p. 138) Thus, a signal degree of progress might well result from a careful consideration of the kinds of methodological issues discussed here. Giere (1988, p. 139) is on the mark when he writes, "Scientists' knowledge of the technology used in experimentation is far more reliable than their knowledge of the subject matter of their experiments." We should attend to knowledge at the level of what Giere calls *embodied knowledge*: embodied in the technology used in performing experiments. By developing the understanding of research methods we should more easily develop a reliable and meaningful base of evidence that will allow us to understand and to explain the complex phenomena encountered when we consider the behavior of organisms in their natural environment. The present article is intended to point the direction for progress in the development of more appropriate methodological approaches to the complex problems that face behavioral science.

Author notes. I thank my colleagues David Funder, Patty O'Neill, and Larry Rosenblum who read drafts of this paper. They helped tame my rhetoric and forced me to demystify some of the logic of my arguments.

References

Anderson, J. R. (1974). Verbatim and propositional representation of sentences in immediate and long-term memory. *Verbal Behavior*, *13*, 149–162.

Anderson, J. R. (1985). *Cognitive psychology and its implications* (2nd ed.). Freeman.

Bakan, D. (1967). The test of significance in psychological research. In D. Bakan (Ed.), *On method: Toward a reconstruction of psychological investigation* (pp. 1–29). Jossey-Bass.

Banaji, M. R., & Crowder, R. G. (1989). The bankruptcy of everyday memory. *American Psychologist*, *44*, 1185–1193.

Bandura, A. (1978). On paradigms and recycled ideologies. *Cognitive Therapy and Research*, *2*, 79–103.

Block, J. (1977). Advancing the psychology of personality: Paradigmatic shift or improving the quality of research? In D. Magnusson & N. S. Endler (Eds.), *Personality at the crossroads: Current issues in interactional psychology* (pp. 37–64). Lawrence Erlbaum Associates.

Bronfenbrenner, U. (1977). Toward an experimental ecology of human development. *American Psychologist*, *32*, 513–531.

Brown, H. I. (1977). *Perception, theory, and commitment*. University of Chicago Press.

Brown, R., & Hanlon, C. (1970). Derivational complexity and order of acquisition in child speech. In J. R. Hayes (Ed.), *Cognition and the development of language* (pp. 11–53). Wiley.

Brunswik, E. (1952). *The conceptual framework of psychology*. University of Chicago Press.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments*. University of California Press.

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, *54*, 297–312.

Campbell, D. T. (1986). Relabeling internal and external validity for applied social scientists. In W. M. K. Trochim (Ed.), *Advances in quasi-experimental design and analysis* (pp. 67–77). Jossey-Bass.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81–105.

Chamberlin, T. C. (1897). Studies for students: The method of multiple working hypotheses. *Journal of Geology*, *5*, 837–848.

Chow, S. L. (1988). Significance test or effect size? *Psychological Bulletin*, *103*, 81–105.

Cochran, W. G., & Cox, G. M. (1957). *Experimental designs* (2nd ed.). Wiley.

Cohen, J. (1962). The statistical power of abnormal-social psychology research: A review. *Journal of Abnormal and Social Psychology*, *65*, 145–163.

Cronbach, L. J. (1986). Social inquiry by and for earthlings. In D. W. Fiske & R. A. Shweder (Eds.), *Metatheory in social science* (pp. 83–107). University of Chicago Press.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281–302.

Epstein, S. (1980). The stability of behavior: II. Implications for psychological research. *American Psychologist*, *35*, 790–806.

Feyerabend, P. (1975). *Against method*. Verso.

Forbes, L. S. (1990). A note on statistical power. *The Auk*, *107*, 438–439.

Funder, D. C. (1987). Errors and mistakes: Evaluating the accuracy of social judgment. *Psychological Bulletin*, *101*, 75–90.

Funder, D. C., & Ozer, D. J. (1983). Behavior as a function of the situation. *Journal of Personality and Social Psychology*, *44*, 107–112.

Gardiner, J. M. (1990). A new psychology of memory? *Contemporary Psychology*, *35*, 215–218.

Ghiselin, M. T. (1971). The individual in the Darwinian revolution. *New Literary History*, *3*, 113–134.

Giere, R. N. (1988). *Explaining science: A cognitive approach*. University of Chicago Press.

Goddard, G. V. (1980). Component properties of the memory machine: Hebb revisited. In P. W. Jusczyck & R. M. Klein (Eds.), *The nature of thought: Essays in honor of D. O. Hebb* (pp. 231–247). Lawrence Erlbaum Associates.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, *82*, 1–20.

Hammond, K. R. (1978). *Psychology's scientific revolution: Is it in danger?* (No. 211). Center for Research on Judgement; Policy.

Hammond, K. R., Hamm, R. M., & Grassia, J. (1986). Generalizing over conditions by combining the multitrait-multimethod matrix and the representative design of experiments. *Psychological Bulletin*, *100*, 257–269.

Harcum, E. R. (1990). Methodological versus empirical literature: Two views on casual acceptance of the null hypothesis. *American Psychologist*, *45*, 404–405.

Hebb, D. O. (1949). *The organization of behavior*. Wiley.

Hochberg, J. (1988). Visual perception. In R. C. Atkinson, R. Herrnstein, G. Lindzey, & R. D. Luce (Eds.), *Stevens' handbook of experimental psychology* (2nd ed., Vol. 1, pp. 295–375). Wiley.

Hull, D. L. (1988). *Science as a process*. University of Chicago Press.

Jöreskog, K. G., & Sörbom, D. (1981). *LISREL v: Analysis of linear structural relationships by maximum likelihood and least squares methods*. National Education Resources.

Kanekar, S. (1990). Statistical significance as a continuum. *American Psychologist*, *45*, 296.

Kolers, P. A. (1979). A pattern-analyzing basis of recognition. In L. S. Cermak & F. I. M. Craik (Eds.), *Levels of processing in human memory* (pp. 363–384). Lawrence Erlbaum Associates.

McGuire, W. J. (1973). The yin and yang of progress in social psychology: Seven koan. *Journal of Personality and Social Psychology*, *26*, 446–456.

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

Meehl, P. E. (1986). What social scientists don't understand. In D. W. Fiske & R. A. Shweder (Eds.), *Metatheory in social science* (pp. 315–338). University of Chicago Press.

Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, *66*, 195–244.

Milner, P. M. (1957). The cell assembly: Mark II. *Psychological Review*, *64*, 242–252.

Mook, D. G. (1983). In defense of external invalidity. *American Psychologist*, *38*, 379–387.

Mook, D. G. (1989). The myth of external validity. In L. W. Poon, D. C. Rubin, & B. A. Wilson (Eds.), *Everyday cognition in adult and late life* (pp. 25–43). Cambridge University Press.

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

Murphy, K. R. (1990). If the null hypothesis is impossible, why test it? *American Psychologist*, *45*, 403–404.

Neisser, U. (1976). *Cognition and reality*. W. H. Freeman.

Neisser, U., & Winograd, E. (1988). *Remembering reconsidered: Ecological and traditional approaches to the study of memory*. Cambridge University Press.

Nunnally, J. (1960). The place of statistics in psychology. *Educational and Psychological Measurement*, *20*, 641–650.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioural sciences*. Wiley.

Parke, R. D. (1976). Social cues, social control, and ecological validity. *Merrill-Palmer Quarterly*, *22*, 111–123.

Petrinovich, L. (1976). Molar reductionism. In L. Petrinovich & J. L. McGaugh (Eds.), *Knowing, thinking, and believing: Festschrift for Professor David Krech* (pp. 11–27). Academic Press.

Petrinovich, L. (1979). Probabilistic functionalism: A conception of research method. *American Psychologist*, *34*, 373–390.

Petrinovich, L. (1981). A method for the study of development. In K. Immelmann, G. Barlow, L. Petrinovich, & M. Main (Eds.), *Behavioral development: The Bielefeld Interdisciplinary Project* (pp. 95–130). Cambridge University Press.

Petrinovich, L. (1989). Representative design and the quality of generalization. In L. W. Poon, D. C. Rubin, & B. A. Wilson (Eds.), *Everyday cognition in adulthood and late life* (pp. 11–24). Cambridge University Press.

Platt, J. R. (1964). Strong inference. *Science*, *146*(3642), 347–353.

Popper, K. R. (1959). *The logic of scientific discovery*. Basic Books.

Posner, M. I. (1969). Abstraction and the process of recognition. In G. H. Bower & J. T. Spence (Eds.), *The psychology of learning and motivation* (pp. 43–100). Academic Press.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*, 638–641.

Rozeboom, W. W. (1960). The fallacy of the null hypothesis significance test. *Psychological Bulletin*, *57*, 416–428.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, *105*, 309–316.

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

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of the American Statistical Association*, *54*, 30–34.

Stokols, D. (1982). Environmental psychology: A coming of age. In A. G. Kraut (Ed.), *The G. Stanley Hall lecture series* (Vol. 2, pp. 159–205). American Psychological Association.

Tukey, J. W. (1969). Analyzing data: Sanctification or detective work? *American Psychologist*, *24*, 83–91.

Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, *20*, 273–281.

Yeaton, W. H., & Sechrest, L. B. (1987). No-difference research. *New Directions for Program Evaluation*, *34*, 67–82.