Philosophy of Science, 1967, Vol. 34, 103–115.

#74

Philosophy of Science

June, 1967

THEORY-TESTING IN PSYCHOLOGY AND PHYSICS: A METHODOLOGICAL PARADOX*

PAUL E. MEEHL[1]

Minnesota Center for Philosophy of Science

Because physical theories typically predict numerical values, an improvement in experimental precision reduces the tolerance range and hence increases corroborability. In most psychological research, improved power of a statistical design leads to a prior probability approaching ½ of finding a significant difference in the theoretically predicted direction. Hence the corroboration yielded by "success" is very weak, and becomes weaker with increased precision. "Statistical significance" plays a logical role in psychology precisely the reverse of its role in physics. This problem is worsened by certain unhealthy tendencies prevalent among psychologists, such as a premium placed on experimental "cuteness" and a free reliance upon *ad hoc* explanations to avoid refuation.

The purpose of the present paper is not so much to propound a doctrine or defend a thesis (especially as I should be surprised if either psychologists or statisticians were to disagree with whatever in the nature of a "thesis" it advances), but to call the attention of logicians and philosophers of science to a puzzling state of affairs in the currently accepted methodology of the behavior sciences which I, a psychologist, have been unable to resolve to my satisfaction. The puzzle, sufficiently striking (when clearly discerned) to be entitled to the designation "paradox," is the following: *In the physical sciences, the usual result of an improvement in experimental design, instrumentation, or numerical mass of data, is to increase the difficulty of the "observational hurdle" which the physical theory of interest must successfully surmount; whereas, in psychology and some of the allied behavior sciences, the usual effect of such improvement in experimental precision is to provide an easier hurdle for the theory to surmount.* Hence what we would normally think of as improvements in our experimental method tend (when predictions materialize) to yield

* Received March, 1967.

[1]I wish to express my indebtedness to Dr. David T. Lykken, conversations with whom have played a major role in stimulating my thinking along these lines, and whose views and examples have no doubt influenced the form of the argument in this paper. For an application of these and allied considerations to a specific example of poor research in psychology, see [7].

stronger corroboration of the theory in physics, since to remain unrefuted the theory must have survived a more difficult test; by contrast, such experimental improvement in psychology typically results in a *weaker* corroboration of the theory, since it has now been required to survive a more lenient test [3] [9] [10].

Although the point I wish to make is one in logic and methodology of science and, as I think, does not presuppose adoption of any of the current controversial viewpoints in technical statistics, a brief exposition of the process of statistical inference as we usually find it in the social sciences is necessary. (The philosopher who is unfamiliar with this subject-matter may be referred to any good standard text on statistics, such as the widely used book by Hays [5] which includes a clear and succinct treatment of the main points I shall briefly summarize here.)

On the basis of a substantive psychological theory T in which he is interested, a psychologist derives (often in a rather loose sense of 'derive') the consequence that an observable variable x will differ as between two groups of subjects, Sometimes, as in most problems of clinical or social psychology, the two groups are defined by a property the individuals under study already possess, e.g., social class, sex, diagnosis, or measured I.Q. Sometimes, as is more likely to be the case in such fields as psychopharmacology or psychology of learning, the contrasted groups are defined by the fact that the experimenter has subjected them to different experimental influences, such as a drug, a reward, or a specific kind of social pressure. Whether the contrasted groups are specified by an "experiment of nature" where the investigator takes the organisms as he finds them, or by a true "experiment" in the more usual sense of the word, is not crucial for the present argument; although, as will be seen, the implications of my puzzle for theory-testing are probably more perilous in the former kind of research than in the latter.

According to the substantive theory T, the two groups are expected to differ on variable x, but it is recognized that errors of (a) measurement and (b) random sampling will, in general, produce *some* observed difference between the averages of the groups studied, even if their total population did not differ in the true value of $\bar{x}$ [= mean of x].

Example: We are interested in the question whether girls are brighter than boys (i.e., that $\mu_g - \mu_b = \delta_{gb} > 0$). We do not have perfectly reliable measures of intelligence, and we are furthermore not in a position to measure the intelligence of all boys and girls in the hypothetical population about which we desire to make a general statement. Instead we must be content with fallible I.Q. scores, and with a sample of school children drawn from the hypothetical population. Each of these sources of error, measurement error and random sampling error, contributes to an untrustworthiness in the computed value we obtain for the average intelligence, $\bar{x}_b$ of the boys and also for $\bar{x}_g$, that of the girls. If we observe a difference of, say d = 5 I.Q. points in a sample of 100 boys and 100 girls, we must have some method to infer whether this obtained observational difference between the two groups reflects a real difference or one which is merely apparent, i.e., due to the combined effect of errors of measurement and sampling. We do this by means of a "statistical significance test," the mathematics of which is not relevant here, except to say that by combining the principles of probability with a characterization of the procedure by which the samples were constituted, and quantifying the variation in observed intelligence score *within* each of the two groups being contrasted, it is possible to

employ a formula which utilizes the observed averages together with the observed variations and sample sizes so as to answer certain relevant kinds of questions. Among such questions is the following: "If there were, in fact, no real difference in average I.Q. between the population of boys and girls, with what relative frequency would an investigator find a difference—in relation to the observed intra-group variation—of the magnitude our observations have actually found"?

The statistical hypothesis, that there is no population difference between boys and girls in I.Q., which is called the "null hypothesis" ($H_0 : \overline{\delta} = 0$) is used to generate a random sampling distribution of the statistic ("*t*-test") employed in testing the presence of a significant difference. If the observed data would be very improbable on the hypothesis that H_0 obtained, we abandon H_0 in favor of its alternative. We conclude that since H_0 is false, its alternative, i.e., that there exists a real average difference between the sexes, obtains. In the past, it was customary to deal with what may be called the "point-null" hypothesis, which says that there is zero difference between the two averages in the populations. In recent years it has been more explicitly recognized that what is of theoretical interest is not the mere presence of *difference* (i.e., that H_0 is false, i.e., that $\mu_b \neq \mu_g$) but rather the presence of a difference *in a certain direction* (in this case, that $\mu_g > \mu_b$). It is therefore increasingly frequent that the behavior scientist employs the so-called "directional null hypothesis," say H_2, instead of the point-null hypothesis H_0. If our substantive theory T involves the prediction that the average I.Q. of girls in the entire population exceeds that of boys, we test the alternative to this statistical hypothesis about the population, i.e., that *either the average I.Q. of boys exceeds that of girls* (H_2) *or that there is no difference* (H_0). That is, we adopt for statistical test (with the anticipation of refuting it) a disjunction of the old-fashioned point-null hypothesis H_0 with the hypothesis H_2 that H_0 is false and it is false in a direction *opposite* to that implied by our substantive theory. However, this directional null hypothesis ($H_{02} : \mu_g \leq \mu_b$), unlike the old-fashioned point-null hypothesis ($H_0 : \mu_g = \mu_b$), does not generate a theoretically expected distribution, because it is not precise, i.e., it does not specify a point-value for the unknown parameter $\overline{\delta} = \mu_{girls} - \mu_{boys}$). However, we can employ it as we do the point-null hypothesis, by reasoning that *if* the point-null hypothesis H_0 obtained in the state of nature, *then* an observed difference (in the direction that our substantive theory predicts) of such-and-such magnitude, has a calculable probability; and that calculable probability is an upper bound upon the desired (but unknown) probability based on $H_{02} : \mu_g \leq \mu_b$. That is to say, if the probability of observed girl-over-boy difference ($d_{gb} = \overline{x}_g - \overline{x}_b$) arising through random error is *p,* given the point-null hypothesis $H_0 : \mu_g = \mu_b$, then the probability of the observed difference arising randomly given any of the point-hypotheses constituting $H_2 : \mu_g < \mu_b$ will of course be less than *p*. Hence *p* is an upper bound on this probability for the inexact directional null hypothesis ($H_{02} : \mu_g \leq \mu_b$). Proceeding in this way directs our interest to only one tail of the theoretical random sampling distribution instead of both tails, which has given rise to a certain amount of controversy among statisticians, but that controversy is not relevant here. (For an excellent clarifying discussion, see Kaiser [6]). Suffice it to say that having formulated a directional null hypothesis H_{02} which is the alternative to the statistical hypothesis of interest H_1, and which includes the point-null hypothesis H_0 as one (very unlikely) possibility for the state of nature,

we then carry out the experiment with the anticipation of *refuting* this directional null hypothesis, thereby confirming the alternative statistical hypothesis of interest (H_1), and, since H_1 in turn was implied by the substantive theory T, of corroborating T.

In such a situation we know in advance that we are in danger of making either of two sorts of "errors," not in the sense of committing scientific mistakes but in the sense of (rationally) inferring what is objectively a false conclusion. If the null hypothesis (point or directional) is in fact true, but due to the combination of measurement and sampling errors we obtain a value which is so improbable upon H_2 or H_0 that we decide in favor of their alternative H_1, we will then have committed what is known as an *error of the first kind* or *Type I Error*. An error of the first kind is a statistical inference that the null hypothesis is false, when in the state of nature it is actually true. This means we will have concluded in favor of a statistical statement H_1 which flowed as a consequence of our substantive theory T, and therefore we will believe ourselves to have obtained empirical support for T, whereas in reality this statistical conclusion is false and, consequently, such support for the substantive theory is objectively lacking. Measurement and sampling error may, of course, also result in a sampling deviation in the opposite direction; or, the true difference $\overline{\delta}$ may be so small that even if our sample values were to coincide exactly with the true ones, the sheer algebra of the significance test would not enable us to reach the prespecified level of statistical significance. If we conclude until further notice that the directional null hypothesis H_{02} is tenable, on the grounds that we have failed to refute it by our investigation, then we have failed to support its statistical alternative H_1, and therefore failed to confirm one of the predictions of the substantive theory T. Retention of the null hypothesis H_{02} when it is in fact false is known as an *error of the second kind* or *Type II Error*.

In the biological and social sciences there has been widespread adoption of the probabilities .01 or .05 as the allowable theoretical frequency of Type I errors. These values are called the 1% and 5% "levels of significance." It is obvious that there is an inverse relationship between the probabilities of the two kinds of errors, so that if we adopt a significance level which increases the frequency of Type I errors, such a policy will lead to a greater number of claims of statistically significant departure from the null hypothesis; and, therefore, in whatever unknown fraction of all experiments performed the null hypothesis is in reality false, we will more often (correctly) conclude its falsity, i.e., we will thereby be reducing the proportion of Type II errors.

Suppose we hold fixed the theoretically calculable incidence of Type I errors. Thus, we determine that *if* the null hypothesis is in fact true in the state of nature, we do not wish to risk erroneously concluding that it is false more than, say, five times in 100. Holding this *5%* significance level fixed (which, as a form of scientific strategy, means leaning over backward not to conclude that a relationship exists when there isn't one, or when there is a relationship in the wrong direction), we can decrease the probability of Type II errors by improving our experiment in certain respects. There are three general ways in which the frequency of Type II errors can be decreased (for fixed Type I error-rate), namely, (a) by improving the logical structure of the experiment, (b) by improving experimental techniques such as the control of extraneous variables which contribute to intragroup variation (and

hence appear in the denominator of the significance test), and (c) by increasing the size of the sample. Given a specified true difference in the range of H_1, the complement (1 – p) of the probability of a Type II error is known as the *power,* and an improvement in the experiment by any or all of these three methods yields an increase in power (or, to use words employed by R. A. Fisher, the experiment's "sensitiveness" or "precision"). For many years relatively little emphasis was put upon the problem of power, but recently this concept has come in for a good deal of attention. Accordingly, up-to-date psychological investigators are normally expected to include some preliminary calculations regarding power in designing their experiments. We select a logical design and choose a sample size such that it can be said in advance that if one is interested in a true difference provided it is at least of a specified magnitude (i.e., if it is smaller than this we are content to miss the opportunity of finding it), the probability is high (say, 80%) that we will successfully refute the null hypothesis. See, for example, Cohen's literature sampling [4] on the problem of power. For an incisive critique of the whole approach, a critique which has been given far less respectful attention than it deserves (conspiracy of silence?), I recommend Rozeboom's excellent contribution [11]. But I should emphasize that my argument in this paper does not hinge upon the reader's agreement with Rozeboom's very strong attack (although I, myself, incline to go along with him).

It is important to keep clear the distinction between the *substantive theory* of interest and the *statistical hypothesis* which is derived from it [2]. In the I.Q. example there was almost no substantive theory or a very impoverished one; i.e., the question being investigated was itself stated as a purely statistical question about the average I.Q. of the two sexes. In the great majority of investigations in psychology the situation is otherwise. Normally, the investigator holds some substantive theory about unconscious mental processes, or physiological or genetic entities, or perceptual structure, or about learning influences in the person's past, or about current social pressures, which contains a great deal more content than the mere statement that the population parameter of an observational variable is greater for one group of individuals than for another. While no competent psychologist is unaware of this obvious distinction between a substantive psychological theory T and a statistical hypothesis H implied by it, in practice there is a tendency to conflate the substantive theory with the statistical hypothesis, thereby illicitly conferring upon T somewhat the same degree of support given H by a successful refutation of the null hypothesis. Hence the investigator, upon finding an observed difference which has an extremely small probability of occurring on the null hypothesis, gleefully records the tiny probability number "$p < .001$," and there is a tendency to feel that the extreme smallness of this probability of a Type I error is somehow transferable to a small probability of "making a theoretical mistake." It is as if, when the observed statistical result would be expected to arise only once in a thousand times through a Type I statistical error given H_{02}, therefore one's substantive theory T, which entails the alternative H_1, has received some sore of direct quantitative support of magnitude around .999 [= 1 – .001].

To believe this literally would, of course, be an undergraduate mistake of which no competent psychologist would be guilty; I only want to point to the fact that there is subtle tendency to "carry over" a very small probability of a Type I error into a sizeable resulting confidence in the truth of the substantive theory, even among in-

vestigators who would never make an explicit identification of the one probability number with the complement of the other.

One reason why the directional null hypothesis ($H_{02} : \mu_g \leq \mu_b$) is the appropriate candidate for experimental refutation is the universal agreement that the old point-null hypothesis ($H_0 : \mu_g = \mu_b$) is [quasi-] always false in biological and social science. Any dependent variable of interest, such as I.Q., or academic achievement, or perceptual speed, or emotional reactivity as measured by skin resistance, or whatever, depends mainly upon a finite number of "strong" variables characteristic of the organisms studied (embodying the accumulated results of their genetic makeup and their learning histories) plus the influences manipulated by the experimenter. Upon some complicated, unknown mathematical function of this finite list of "important" determiners is then superimposed an indefinitely large number of essentially "random" factors which contribute to the intragroup variation and therefore boost the error term of the statistical significance test. In order for two groups which differ in some identified properties (such as social class, intelligence, diagnosis, racial or religious background) to differ not at all in the "output" variable of interest, it would be necessary that all determiners of the output variable have precisely the same average values in both groups, or else that their values should differ by a *pattern of amounts of difference* which precisely counterbalance one another to yield a net difference of zero. Now our general background knowledge in the social sciences, or, for that matter, even "common sense" considerations, makes such an exact equality of all determining variables, or a precise "accidental" counterbalanceing of them, so extremely unlikely that no psychologist or statistician would assign more than a negligibly small probability to such a state of affairs.

Example: Suppose we are studying a simple perceptual-verbal task like rate of color-naming in school children, and the independent variable is father's religious preference. Superficial consideration might suggest that these two variables would not be related, but a little thought leads one to conclude that they will almost certainly be related by *some* amount, however small. Consider, for instance, that a child's reaction to any sort of school-context task will be to some extent dependent upon his social class, since the desire to please academic personnel and the desire to achieve at a performance (just because it is a *task,* regardless of its intrinsic interest) are both related to the kinds of sub-cultural and personality traits in the parents that lead to upward mobility, economic success, the gaining of further education, and the like. Again, since there is known to be a sex difference in color-naming, it is likely that fathers who have entered occupations more attractive to "feminine" males will (on the average) provide a somewhat more feminine father-figure for identification on the part of their male offspring, and that a more refined color vocabulary, making closer discriminations between similar hues, will be characteristic of the ordinary language of such a household. Further, it is known that there is a correlation between a child's general intelligence and its father's occupation, and of course there will be *some* relation, even though it may be small, between a child's general intelligence and his color vocabulary, arising from the fact that *vocabulary in general* is heavily saturated with the general intelligence factor. Since religious preference is a correlate of social class, all of these social class factors, as well as the intelligence variable, would tend to influence color-naming performance. Or consider a more extreme and faint kind of relationship. It is quite conceivable that

a child who belongs to a more liturgical religious denomination would be somewhat more color-oriented than a child for whom bright colors were not associated with the religious life. Everyone familiar with psychological research knows that numerous "puzzling, unexpected" correlations pop up all the time, and that it requires only a moderate amount of motivation-plus-ingenuity to construct very plausible alternative theoretical explanations for them.

These armchair considerations are borne out by the finding that in psychological and sociological investigations involving very large numbers of subjects, it is regularly found that almost all correlations or differences between means are statistically significant. See, for example, the papers by Bakan [1] and Nunnally [8]. Data currently being analyzed by Dr. David Lykken and myself, derived from a huge sample of over 55,000 Minnesota high school seniors, reveal statistically significant relationships in 91% of pairwise associations among a congeries of 45 miscellaneous variables such as sex, birth order, religious preference, number of siblings, vocational choice, club membership, college choice, mother's education, dancing, interest in woodworking, liking for school, and the like. The 9% of non-significant associations are heavily concentrated among a small minority of variables having dubious reliability, or involving arbitrary groupings of non-homogeneous or non-monotonic sub-categories. The majority of variables exhibited significant relationships *with all but three of the others,* often at a very high confidence level ($p < 10^{-6}$).

This line of reasoning is perhaps not quite as convincing in the case of true *experiments,* where the subjects are randomly assigned by the investigator to different experimental manipulations. If the reader is disinclined to follow me here, my overall argument will, for him, be applicable to those kinds of research in social science which study the correlational relationships or group differences between subjects "as they come," but not to the type of investigation which constitutes an experiment in the usual scientific sense. However, I myself believe that even in the strict sense of 'experiment,' the argument is still strong, although the quantitative departures from the point-null H_0 would be expected to run considerably lower on the average. Considering the fact that "everything in the brain is connected with everything else," and that there exist several "general state-variables" (such as arousal, attention, anxiety, and the like) which are known to be at least *slightly* influenceable by practically any kind of stimulus input, it is highly unlikely that *any* psychologically discriminable stimulation which we apply to an experimental subject would exert literally *zero* effect upon any aspect of his performance. The psychological literature abounds with examples of small but detectable influences of this kind. Thus it is known that if a subject memorizes a list of nonsense syllables in the presence of a faint odor of peppermint, his recall will be facilitated by the presence of that odor. Or, again, we know that individuals solving intellectual problems in a "messy" room do not perform quite as well as individuals working in a neat, well-ordered surround. Again, cognitive processes undergo a detectable facilitation when the thinking subject is concurrently performing the irrelevant, non-cognitive task of squeezing a hand dynamometer. It would require considerable ingenuity to concoct experimental manipulations, except the most minimal and trivial (such as a very slight modification in the word order of instructions given a subject) where one could have confidence that the manipulation would be utterly without effect upon the subject's motivational level, attention, arousal, fear of failure,

achievement drive, desire to please the experimenter, distraction, social fear, etc., etc. So that, for example, while there is no very "interesting" psychological theory that links hunger drive with color-naming ability, I myself would confidently predict a significant difference in color-naming ability between persons tested after a full meal and persons who had not eaten for 10 hours, provided the sample size were sufficiently large and the color-naming measurements sufficiently reliable, since one of the effects of the increased hunger drive is heightened "arousal," and anything which heightens arousal would be expected to affect a perceptual-cognitive performance like color-naming. Suffice it to say that there are very good reasons for expecting at least *some* slight influence of almost any experimental manipulation which would differ sufficiently in its form and content from the manipulation imposed upon a control group to be included in an experiment in the first place. In what follows I shall therefore assume that the point-null hypothesis H_0 is, in psychology, [quasi-] always false.

Let us now conceive of a large "theoretical urn" containing counters designating the indefinitely large class of actual and possible substantive theories concerning a certain domain of psychology (e.g., mammalian instrumental learning). Let us conceive of a second urn, the "experimental-design" urn, containing counters designating the indefinitely large set of possible experimental situations which the ingenuity of man could devise. (If anyone should object to my conceptualizing, for purposes of methodological analysis, such a heterogeneous class of theories or experiments, I need only remind him that such a class is universally presupposed in the logic of statistical significance testing.) Since the point-null hypothesis H_0 is [quasi-] always false, almost every one of these experimental situations involves a non-zero difference on its output variable (parameter). Whichever group we (arbitrarily) designate as the "experimental" group and the "control" group, in half o£ these experimental settings the true value of the dependent variable difference (experimental minus control) will be positive, and in the other half negative.

It may be objected that this is a use of the Principle of Insufficient Reason and presupposes one particular answer to some disputed questions in statistical theory (as between the Bayesians and the Fisherians). But I must emphasize that I have said nothing about the *form* or *range* or other parametric characteristics of the distribution of true differences. I have merely said that the point-null hypothesis H_0 is always false, and I have then *assigned,* in a strictly random fashion, the names "experimental" and "control" to the two groups which a given experimental setup treats in two different ways. That is, it makes no difference here whether a group of "subjects learning nonsense syllables while squeezing a hand dynamometer is called the experimental group, or whether we call "experimental" the group that learns the nonsense syllables without such squeezing. Hence my use of the Principle of Insufficient Reason is one of those legitimate, non-controversial uses following directly when the basic principles of probability are applied to a specification of procedure for random assignment.

We now perform a random pairing of the counters from the "theory" urn with the counters from the "experimental" urn, and arbitrarily stipulate—quite irrationally—that a "successful" outcome of the experiment means that the difference favors the experimental group $[\mu_E - \mu_C > 0]$, This preposterous model, which is of course much worse than anything that can exist even in the most primitive of the social sciences, provides us with a lower bound for the expected frequency of a

theory's successfully predicting the direction in which the null hypothesis fails, *in the state of nature* (i.e., we are here not considering sampling problems, and therefore we neglect errors of either the first or the second kind). It is obvious that if the point-null hypothesis H_0 is [quasi-]always false, and there is no logical connection between our theories and the direction of the experimental outcomes, then if we arbitrarily assign one of the two directional hypotheses H_1 or H_2 to each theory, that hypothesis will be correct half of the time, i.e., in half of the arbitrary urn-counter-pairings. Since even my late, uneducated grandmother's common-sense psychological theories had non-zero verisimilitude; we can safely say that the value $p = ½$ is a lower bound on the success-frequency of experimental "tests," assuming our experimental design had perfect power.

Countervailing the unknown increment over $p = ½$ which arises from the fact that the experimental and theoretical counters are not thus drawn randomly (since our theories do possess, on the average, at least some tiny amount of verisimilitude), there is the statistical factor that among the counter-pairings which are accidentally "successful" (in the sense that the state of nature falsifies the null hypothesis in the expected direction), we will sometimes fail to refute it because of measurement and sampling errors, since our experiments will always, in practice, have less than perfect power. Even though the point-null hypothesis H_0 is always false, so that the directional null hypothesis H_{02} is false in the (theoretically pseudo-predicted) direction half the time, we will sometimes fail to discover this because of Type II errors. Without making illegitimate prior-probability assumptions concerning the actual distribution of true differences in the whole vast world of psychological experimental contexts, one cannot say anything definite about the extent to which this countervailing influence of Type II errors will wash out (or even overcome) the fact that our theories tend to have non-negligible verisimilitude. But by setting aside this latter fact, i.e., by assuming counterfactually that there is *no connection whatever between our theories and our experimental designs* (the two-urn idealization), thereby fixing the expected frequency of successful refutations of the directional null hypothesis H_{02} at $p = ½$ for experiments of *perfect power;* it follows that as the power of our experimental designs and significance tests is increased by any of the three methods described above, we approach $p = ½$ as the limit of our expected frequency of "successful outcomes," i.e., of attaining statistically significant experimental results in the theoretically predicted direction.

I conclude that the effect of increased precision, whether achieved by improved instrumentation and control, greater sensitivity in the logical structure of the experiment, or increasing the number of observations, is to yield a probability approaching ½ of corroborating our substantive theory by a significance test, *even if the theory is totally without merit.* That is to say, the ordinary result of improving our experimental methods and increasing our sample size, proceeding in accordance with the traditionally accepted method of theory-testing by refuting a directional null hypothesis, yields a prior probability $p \simeq ½$ and very likely somewhat above that value by an unknown amount. It goes without saying that successfully negotiating an experimental hurdle of this sort can constitute only an extremely weak corroboration of any substantive theory, *quite apart from currently disputed issues of the Bayesian type regarding the assignment of prior probabilities to the theory itself.*

So far as I am able to discern, this methodological truth is either unknown or

systematically ignored by most behavior scientists. I do not know to what extent this is attributable to confusion between the substantive theory T and the statistical hypothesis H_1, with the resulting mis-assignment of the probability $(1 - p_I)$ complementary to the significance level p_I attained, to the "probability" of the substantive theory; or to what extent it arises from insufficient attention to the truism that the point-null hypothesis H_0 is [quasi-]always false. It seems unlikely that most social science investigators would think in their usual way about a theory in meteorology which "successfully predicted" that it would rain on the 17th of April, given the antecedent information that it rains (on the average) during half the days in the month of April!

But this is not the worst of the story. Inadequate appreciation of the extreme weakness of the test to which a substantive theory T is subjected by merely predicting a directional statistical difference $\bar{d} > 0$ is then compounded by a truly remarkable failure to recognize the logical asymmetry between, on the one hand, (formally invalid) "confirmation" of a theory via affirming the consequent in an argument of form: [$T \supset H_1$, H_1, infer T], and on the other hand the deductively tight *refutation* of the theory *modus tollens* by a falsified prediction, the logical form being: [$T \supset H_1$, $\sim H_1$, infer $\sim T$].

While my own philosophical predilections are somewhat Popperian, I daresay any reader will agree that no full-fledged Popperian philosophy of science is presupposed in what I have just said. The destruction of a theory *modus tollens* is, after all, a matter of deductive logic; whereas that the "confirmation" of a theory by its making successful predictions involves a much weaker kind of inference. This much would be conceded by even the most anti-Popperian "inductivist." The writing of behavior scientists often reads as though they assumed—what it is hard to believe anyone would explicitly assert if challenged—that successful and unsuccessful predictions are practically on all fours in arguing for and against a substantive theory. Many experimental articles in the behavioral sciences, and, even more strangely, review articles which purport to survey the current status of a particular theory in the light of all available evidence, treat the confirming instances and the disconfirming instances with equal methodological respect, as if one could, so to speak, "Count noses," so that if a theorv has somewhat more confirming than disconfirming instances, it is in pretty good shape evidentially. Since we know that this is already grossly incorrect on purely formal grounds, it is a mistake *a fortiori* when the so-called "confirming instances" have themselves a prior probability, as argued above, somewhere in the neighborhood of ½, quite apart from any theoretical considerations.

Contrast this bizarre state of affairs with the state of affairs in physics. While there are of course a few exceptions, the usual situation in the experimental testing of a physical theory at least involves the prediction of a *form* of function (with parameters to be fitted); or, more commonly, the prediction of a quantitative magnitude (point-value). Improvements in the accuracy of determining this experimental function-form or point-value, whether by better instrumentation for control and making observations, or by the gathering of a larger number of measurements, has the effect of *narrowing* the band of tolerance about the theoretically predicted value. What does this mean in terms of the significance-testing model? It means: *In physics, that which corresponds, in the logical structure of statistical inference, to the old-fashioned point-null hypothesis H_0 is the value which flows as a conse-*

quence of the substantive theory T; so that an increase in what the statistician would call "power" or "precision" has the methodological effect of stiffening the experimental test, of setting up a more difficult observational hurdle for the theory T to surmount. Hence, in physics the effect of improving precision or power is that of *decreasing* the prior probability of a successful experimental outcome if the theory lacks verisimilitude, that is, precisely the reverse of the situation obtaining in the social sciences.

As techniques of control and measurement improve or the number of observations increases, the methodological effect in physics is that a successful passing of the hurdle will mean a greater increment in corroboration of the substantive theory; whereas in psychology, comparable improvements at the experimental level result in an empirical test which can provide only a progressively weaker corroboration of the substantive theory.

In physics, the substantive theory predicts a point-value, and when physicists employ "significance tests," their mode of employment is to compare the theoretically predicted value x_o with the observed mean $\bar{x}_o$, asking whether they differ (in either direction!) by more than the "probable error" of determination of the latter. Hence H : $H_0 = \mu_x$ functions as a point-null hypothesis, and the prior (logical, antecedent) probability of its being correct in the absence of theory approximates zero. As the experimental error associated with our determination of $\bar{x}_o$ shrinks, values of $\bar{x}_o$ consistent with x_0 (and hence, compatible with its implicans T) must lie within a narrow range. In the limit (zero probable error, corresponding to "perfect power" in the significant test) any non-zero difference ($\bar{x}_o - x_o$) provides a *modus tollens* refutation of T. If the theory has negligible verisimilitude, the logical probability of its surviving such a test is negligible. Whereas in psychology, the result of perfect power (i.e., *certain* detection of any non-zero difference in the predicted direction) is to yield a prior probability $p = ½$ of getting experimental results compatible with T, because perfect power would mean guaranteed detection of whatever difference exists; and a difference [quasi] always exists, being in the "theoretically expected direction" half the time if our substantive theories were all of negligible verisimilitude (two-urn model).

This methodological paradox would exist for the psychologist even if he played his own statistical game fairly. The reason for its existence is obvious, namely, that most psychological theories, especially in the so-called "soft" fields such as social and personality psychology, are not quantitatively developed to the extent of being able to generate point-predictions. In this respect, then, although this state of affairs is surely unsatisfactory from the methodological point of view, and stands in great need of clarification (and, hopefully, of constructive suggestions for improving it) from logicians and philosophers of science, one might say that it is "nobody's fault," it being difficult to see just how the behavior scientist could extricate himself from this dilemma without making unrealistic attempts at the premature construction of theories which are sufficiently quantified to generate point-predictions for refutation.

However, there are five social forces and intellectual traditions at work in the behavior sciences which make the research consequences of this situation even worse than they may have to be, considering the state of our knowledge. In addition to (a) failure to recognize the marked evidential asymmetry between confirmation and *modus tollens* refutation of theories, and (b) inadequate appreciation of the extreme weakness of the hurdle provided by the mere directional significance

test, there exists among psychologists (c) a fairly widespread tendency to report experimental findings with a liberal use of *ad hoc* explanations for those that didn't "pan out." This last methodological sin is especially tempting in the "soft" fields of (personality and social) psychology, where the profession highly rewards a kind of "cuteness" or "cleverness" in experimental design, such as a hitherto untried method for inducing a desired emotional state, or a particularly "subtle" gimmick for detecting its influence upon behavioral output. The methodological price paid for this highly-valued "cuteness" is, of course, (d) an unusual ease of escape from *modus tollens* refutation. For, the logical structure of the "cute" component typically involves use of complex and rather dubious auxiliary assumptions, which are required to mediate the original prediction and are therefore readily available as (genuinely) plausible "outs" when the prediction fails. It is not unusual that (e) this *ad hoc* challenging of auxiliary hypotheses is repeated in the course of a series of related experiments, in which the auxiliary hypothesis involved in Experiment 1 (and challenged *ad hoc* in order to avoid the latter's *modus tollens* impact on the theory) becomes the focus of interest in Experiment 2, which in turn utilizes further plausible but easily challenged auxiliary hypotheses, and so forth. In this fashion a zealous and clever investigator can slowly wend his way through a tenuous nomological network, performing a long series of related experiments which appear to the uncritical reader as a fine example of "an integrated research program," *without ever once refuting or corroborating so much as a single strand of the network.* Some of the more horrible examples of this process would require the combined analytic and reconstructive efforts of Carnap, Hempel, and Popper to unscramble the logical relationships of theories and hypotheses to evidence. Meanwhile our eager-beaver researcher, undismayed by logic-of-science considerations and relying blissfully on the "exactitude" of modem statistical hypothesis-testing, has produced a long publication list and been promoted to a full professorship. In terms of his contribution to the enduring body of psychological knowledge, he has done hardly anything. His true position is that of a potent-but-sterile intellectual rake, who leaves in his merry path a long train of ravished maidens but no viable scientific offspring.[2]

Detailed elaboration of the intellectual vices (a)–(e) and their scientific consequences must be left for another place, as must constructive suggestions for how the behavior scientist can improve his situation. My main aim here has been to call

[2] Since the readers of this journal cannot, by and large, be expected to possess familiarity with the field of psychology or the contributions of various psychologists, and since quantitative empirical documentation of these admittedly impressionistic comments is still in preparation for subsequent presentation elsewhere, it is perhaps neither irrelevant nor in bad taste to present a few biographical data. Lest the philosophical reader wonder (quite appropriately) whether these impressions of the psychological literature ought perhaps to be dismissed as mere "sour grapes" from an embittered, low-publication psychologist *manqué,* it may be stated that the author (a past president of the American Psychological Association) has published over 70 technical books or articles in both "hard" and "soft" fields of psychology, is a recipient of the Association's Distinguished Scientific Contributor Award, also of the Distinguished Contributor Award of the Division of Clinical Psychology, has been elected to Fellowship in the American Academy of Arts and Sciences, and is actively engaged in both theoretical and empirical research at the present time. He's not mad at anybody—but he is a bit distressed at the state of psychology.

the attention of logicians and philosophers of science to what, as I think, is an important and difficult problem for psychology, or for any science which is largely in a primitive stage of development such that its theories do not give rise to point-predictions.

REFERENCES

[I] Bakan, David, "The test of significance in psychological research," *Psychological Bulletin*, Vol. 66 (1966), pp. 423-437.

[2] Bolles, Robert C., "The difference between statistical hypotheses and scientific hypotheses," *Psychological Reports,* Vol. 11 (1962), pp. 639-645.

[3] Bunge, Mario (ed.). *The critical approach to science and philosophy: essays in honor of Karl R. Popper,* New York: Free Press of Glencoe, Inc., 1964.

[4] Cohen, Jacob, "The statistical power of abnormal-social psychological research: a review," *Journal of abnormal and social Psychology,* Vol. 65 (1962), pp. 145-153.

[5] Hays, William L., *Statistics* for *psychologists,* New York: Holt, Rinehart, and Winston, 1963.

[6] Kaiser, Henry F., "Directional statistical decisions," *Psychological Review,* Vol. 67 (1960), pp. 160-167.

[7] Lykken, David T., "Statistical significance in psychiatric research," *Reports from the Research Laboratories of the Department of Psychiatry, University of Minnesota. Report No.* PR-66-9, Minneapolis: December 30, 1966.

[8] Nunnally, Jum C., "The place of statistics in psychology," *Educational and Psychological* Measurement, Vol. 20 (1960), pp. 641-650.

[9] Popper, Karl R., *The logic of scientific discovery.* New York: Basic Books, 1959.

[10] Popper, Karl R., *Conjectures and refutations,* New York: Basic Books, 1962,

[11] Rozeboom, William W., "The fallacy of the null-hypothesis significance test," *Psychological Bulletin,* Vol. 67 (1960), pp. 416-428.

ljy scan 10/02