DISCUSSION
THEORY CONFIRMATION IN PSYCHOLOGY*

CHRIS SWOYER AND THOMAS C. MONSON†

Universities of Oklahoma and Minnesota

In [13], Meehl drew attention to what he saw as a substantive problem involved in attempts to confirm theories in psychological research, suggesting that there is a radical asymmetry between the manner in which theories in physics and theories in psychology are confirmed. In the physical sciences, improvement in experimental design, instrumentation, or amount of data—in short, increased power—typically makes it more difficult to confirm a theory; in current psychological research the situation is just the opposite.

Recently, however, Herbert Keuth has responded that at least one of Meehl's central contentions is without foundation ([9], p. 546). As a result of Keuth's paper, philosophers unacquainted with current work in psychology may be left with the mistaken impression that all is well in that domain, or at least, that there exist no substantive problems concerning theory confirmation in the social sciences over and above the well-known infirmities besetting theory confirmation in hard sciences such as physics. We will examine Keuth's criticisms and show why they fail. We will then attempt to clarify the problem of null hypothesis testing and its role in theory confirmation in psychological research and, finally, conclude by suggesting some possible remedial actions.

We plan to take up Keuth's points in order, but a brief restatement of the logic behind the use of significance testing may make what follows clearer to those unacquainted with the contemporary psychological scene. (There is not space here to present anything like a comprehensve discussion of the topic; for that the reader should consult almost any text on statistics, e.g., [6].)[1]

One of the major assumptions made by Meehl is that substantive theories in the physical sciences tend to imply or predict absolute point values whereas substantive theories in psychology tend only to predict directional differences on some variable between two groups of subjects. Of course, theories in physics do not by necessity make point predictions nor do those in psychology necessarily make directional predictions. However, due to less sophisticated

*Received October, 1974.

†We are indebted to Paul Meehl for numerous discussions on these and other topics. He is, however, not responsible for any errors contained in this paper.

[1] To facilitate discussion, we will begin by speaking rather loosely of "rejecting," "accepting" and "confirming" hypotheses; to do full justice to the complex issues here involved would take us much too far afield without substantially contributing to the points under discussion.

Philosophy of Science, 42 (1975) pp. 487–502.

theorizing and measurement, the frequency of directional predictions or hypotheses is much greater in psychology. Since Meehl's grievance is concerned with the difference between null hypothesis testing involving directional predictions and parameter estimation involving point predictions, rather than with a simple distinction between physics and psychology ([13], pp. 112–113), our discussion of examples may be simplified by keeping them in the same domain.

Let us assume that psychologist A develops a theory T_a that relates certain hormonal balances in prospective mothers during pregnancy to an increased likelihood that their babies will be boys rather than girls. Furthermore, assume that these hormonal balances have previously been related to the consumption of *prunectin*, one of the substances found in prunes. Based on some fairly simple reasoning, psychologist A derives a directional hypothesis involving a mean (μ) difference between his two groups (H_a: $\mu_{\text{hi prune juice}} > \mu_{\text{lo prune juice}}$). That is, mothers who consume relatively large amounts of prune juice will have a greater percentage of boys in their families than will mothers who consume relatively little prune juice.

To test H_a, psychologist A asks a sample of mothers to reveal their estimated consumption of prune juice and the percentage of boys in their families. After performing a median split to divide the mothers into those who consume more and those who consume less than an average amount of prune juice, he compares the relative percentage of boys in the two groups.

Traditionally, psychological theories were tested by attempting to refute a simple point-null hypothesis (of the form, H_0: $\mu_1 = \mu_2$), but such an approach has been largely supervened by testing point-directional hypotheses (H_{01}: $\mu_1 \geq \mu_2$ or H_{02}: $\mu_1 \leq \mu_2$). Thus, by refuting a hypothesis like H_{01} we may be thought to confirm, to some degree, a directional hypothesis H_1: $\mu_1 < \mu_2$—a form taken by our example H_a above. Without going into a detailed discussion of statistics (see [6] or, more briefly, [13]), suffice it to say that more than a mere difference in the right direction between the means of the two groups is regarded as necessary to allow us to accept a hypothesis like H_a. Because of possible errors of measurement and random sampling, the variance between the two groups must be sufficiently greater than the variance within groups to reduce the probability of falsely accepting H_a (by falsely rejecting H_{0a}). Tradition has dictated that the long run frequency or probability of these errors should be kept equal to or less than five in a hundred; this type of error is called a *Type I error* and is represented by what is known as the "*p* value." If our *p* value is below some time-honored value such as .05, we will be said to have achieved *statistically significant* results. In current psychological practice such results are crucial in indicating what is thought to indicate a successful experimental outcome, and they play a necessary, if not sufficient, role in determining which papers will be accepted by journals.

If other factors are held constant, the probability of Type I errors can be reduced by improving the logical structure of experiments, improving

experimental techniques, or by increasing sample size. For instance, assume that psychologist A finds that boys constitute, on the average, 75% of the families in which the mothers consume large quantities of prune juice, while boys constitute an average of only 50% of the families in which the mothers drink relatively small amounts of prune juice. If the individual scores are distributed in such a way that rejection of H_{0a} (acceptance of H_a) results in a p value of .50 (fifty-fifty chance of error), then merely by having ten times as many subjects with the same relative configuration of scores (same respective sample variances) within groups and having the same means, we would obtain a p value of less than .05. With one hundred times as many subjects, the p value would be considerably less than .001. Clearly, an increase in sample size, will, *ceteris paribus*, make statistical significance easier to achieve.

Another possible error, a *Type II error*, is the probability or frequency of accepting a null hypothesis (or at least tentatively not rejecting it) when the H_0 is in fact false. The frequency of Type II errors can also be reduced by improving the logical structure of experiments, by improving experimental technique, and by increasing the size of the sample. If the desired probability of Type I errors is established at .05, then increased sample size results in a smaller difference between means being required to reach statistical significance (the rejection of H_{0a} and the acceptance of H_a). For instance, again assume that psychologist A finds that boys constitute an average of 75% of the families with mothers with high prune juice consumption whereas boys constitute an average of 50% of the families with mothers with low prune juice consumption, a difference of 25% between the two groups. But this time assume that the individual scores are distributed in such a way that the p value for rejecting the H_{0a} is .05. With ten times as many subjects and with the same respective sample variances, a mean difference of only 8% between the groups would be significant with the p value less than .05 (approximately $25\%/\sqrt{10}$); with ten thousand times as many subjects, a mean difference of .25% would be sufficient. As the sample size approaches infinity or the population size, the difference between the two means which is required to reach statistical significance approaches zero. In other words, as precision or power is increased, there is a greater range of possible experimental outcomes which would be considered to be consistent with the substantive theory T_a.

In contrast, let us assume that psychologist B develops a theory T_b that also relates certain hormonal balances during pregnancy to relative frequencies of male and female babies, but because of more sophisticated theorizing derives more precise predictions. One of these predictions is H_b: $\mu_{\text{hi prune juice}} - \mu_{\text{lo prune juice}} = 25\%$, the hypothesis that boys will constitute approximately 25% more of families in which mothers consume large amounts of prune juice than they will of families in which mothers consume relatively little prune juice.

To test H_b, psychologist B conducts a research project identical to that

of psychologist A's. In contrast to psychologist A's use of statistical tests involving directional hypotheses, psychologist B decides to examine the 95% confidence interval on the difference between his sample means to determine if his prediction of a 25% difference is included within two standard deviations of the mean sample difference. Since psychologist B is aware that 95% confidence intervals on sample means are supposed to include the population mean an average of ninety-five times out of a hundred, he concludes that this method will afford a good way to ascertain whether or not his H_b will still be a plausible hypothesis. Furthermore, psychologist B realizes that if the 95% confidence interval does not contain his hypothesized value, then he should probably reject H_b since this would be the correct conclusion ninety-five times out of a hundred.

As with directional hypothesis tests, the probability of Type I errors or Type II errors can be reduced by improving the logical structure of an experiment, by improving experimental techniques, or by increasing the size of the sample. However, in contrast to directional hypothesis testing where increased power or precision results in a greater range of possible experimental outcomes consonant with the substantive theory of interest, the opposite will generally occur when point predictions are tested by estimation procedures such as confidence intervals.

For instance, let us assume that like psychologist A, psychologist B finds that boys constitute 25% more of the families in which the mothers consume large quantities of prune juice than they do of those in which the mothers consume relatively little prune juice. If the individual scores are distributed just as they were in psychologist A's research where they result in a p value of .50, then the 95% confidence interval calculated from the data will be $-50\% < \mu_{\text{hi prune juice}} - \mu_{\text{lo prune juice}} < +100\%$. Although this interval contains psychologist B's H_b of a difference of 25%, the size of the interval cannot be taken as very strong evidence for H_b, let alone T_b. However, with one hundred times as many subjects, and with the same sample variances, the 95% confidence interval will be reduced to a much smaller interval ($17.5\% < \mu_{\text{hi prune juice}} - \mu_{\text{lo prune juice}} < 32.5\%$). With ten thousand times as many subjects, the confidence interval will be further reduced to ($24.25\% < \mu_{\text{hi prune juice}} - \mu_{\text{lo prune juice}} < 25.75\%$).

Thus, when point predictions such as H_b are derived from substantive theories such as T_b, the theory must survive a much more difficult test to remain unrefuted in the face of great power and precision. For very high sample numbers, with power approaching 1 (such as the last example involving psychologist B), the hypothesized point value must be well within 1% of the sample mean (or less than 1/100 of the possible range) for the hypothesis, and thus the theory, to remain plausible possibilities. The smallness of the range of possible confirming outcomes should decrease the likelihood that H_b is true if T_b is not true as well.

On the other hand, when directional hypotheses such as H_a are derived from substantive theories such as T_a, the theory must survive a much more

lenient test to remain unrefuted in the face of greater power and precision. With power approaching 1, as a result of extremely large sample sizes (such as the last example concerning psychologist A), the sample mean of the group predicted to be larger must be only negligibly higher than the sample mean of the group predicted to be smaller for the theory to remain a plausible possibility. This results in a state of affairs in which any experimental outcome greater than 0% up to and including 100% in the predicted direction is taken as confirmation of T_a. Since this is approximately half of the range of possible outcomes, it appears that H_a has almost a 50% chance of being accepted even if T_a lacks any verisimilitude. This range, being approximately one hundred times as large as the range acceptable for H_b, could be attributed to many more possible alternative theories (e.g., constipation due to less consumption of prune juice may lead to certain abdominal pressures which might affect the reproduction system, etc.) than could more specific hypotheses such as H_b which limit the number of possible and plausible theories.

In short, with standard directional hypotheses psychologists typically regard a statistically significant outcome as warranting belief that a predicted hypothesis is very likely true and hence that it confirms their theory. (Though, as often interpreted, the logic of the tests does not allow this, this almost always is what happens.) Meehl contends that a "paradox" exists because in the physical sciences, where point predictions abound, greater power and precision produce more novel hypotheses (due to restriction of range) which result in the falsifiability of a greater number of alternative theories. By contrast, in psychology, where directional hypotheses abound, greater power and precision produce less novel hypotheses (due to expansion of range) which result in the falsifiability of a fewer number of alternative theories.[2]

With this background let us turn to Keuth's paper. Employing a distinction of his own devising, Keuth tells us that Meehl presents his methodological paradox in both a "weaker" and a "stronger" version. The weaker version is characterized by Keuth thus:

> In physics T implies H_0. Therefore increasing precision or power of the test will lead to decreasing probability of accepting H_0 and T tentatively along with it, given that T lacks versimilitude. In the social sciences the situation is precisely the reverse. ([9], p. 538; quoted from [13], p. 113)

Keuth notes that 'verisimilitude' is not a technical term of probability theory and that Popper, who introduced the term into current discussions of philosophy of science, has not presented a "tenable definition." This is correct, although of course there are no definitions which would likely be regarded as tenable

[2]The magnitude of the problem is compounded by such common, but false, assumptions as that experimental outcomes with large sample size are theoretically preferable to experimental outcomes with small sample size even when the p values are equivalent and are tested by directional hypotheses. (A number of professional psychologists actually made this assumption [17], though they are not alone—so did a number of dentists [2].)

by a majority of philosophers of even such important terms as 'true', 'random', 'probability', and the like. Lack of definition does not mean that a term cannot be intelligently and profitably employed, especially in a discussion which, as Meehl's does, moves at an intuitive level. Indeed, Meehl's points are intended to hold regardless of the specific definitions one might accept of 'true' or 'probable'. Moreover, inasmuch as Popper does provide a reasonably clear and even technical presentation of what he means by verisimilitude (e.g. [15], pp. 228–234; 391–402; and *passim*), the term is not totally obscure.

But despite Meehl's rather straightforward use of this term, Keuth says of Meehl's "weak" version:

> We may, however, easily interpret Meehl's statement [as paraphrased by Keuth above] as saying that increasing precision will lead to decreasing probability of accepting H_0, *given that it is false* ([9], p. 539; italics ours).

But to say that a theory lacks verisimilitude is not to say that H_0 is false! One might, of course, be led to assert this as a result of confusing substantive scientific theories with statistical hypotheses. However, this should not be construed as a confusion precipitated by unclear statements of Meehl's. Meehl's statement that "no competent psychologist is unaware of this obvious distinction between a substantive psychological theory T and a statistical hypothesis H implied by it" ([13], p. 107) and his reference to the article by Bolles [3], should have obviated the confusion, however one would fill in the details concerning the distinction. Meehl is readily aware that increasing precision in hypothesis testing will lead to decreasing probability of accepting H_0, *given that H_0 is false;* this is inherent in the statistical definition of power and Meehl would not have had any reason to write his paper if that is what he meant! What Meehl was attempting to convey is that in psychology, with increasing precision or power, there is a decreasing probability of accepting the point null hypothesis H_0 and an increasing probability of accepting the statistical hypothesis of relevance, even if the *substantive theory* (not the *statistical hypothesis*) is false.

Keuth's discussion of the "stronger" version demonstrates the same confusion:

> [Meehl] obviously means that the probability of accepting H_0 (and T tentatively along with it) given H_0 is false, approaches *zero*, if the precision of the test grows perfect . . . Meehl claims that the probability of accepting a *directional* hypothesis H_{II}, "*even if the theory* [which implies the hypothesis] *is totally without merit,*" approaches $p = 0.5$, when the power grows perfect . . . If the level of significance is, as usual, set at 0.01 or 0.05, the probability of accepting H_{II}, given it is false, cannot be $p = 0.5$. ([9], p. 539)

But the point is that H_{II} can be true without the substantive theory T being

true. Meehl is concerned with the case where the probability supposedly approaches 0.5 of accepting H_{II}, given that $T(not\ H_{II})$ is false or "lacks verisimilitude." That Keuth seems to overlook the distinction between theories and hypotheses which they imply is further evinced by his remarks that there is nothing wrong with increasing the chances of accepting a true hypothesis and his rhetorical questioning of whether or not we should view such circumstances as involving a more lenient test ([9], p. 542). Meehl argues that it does involve a more lenient test, *not of the hypothesis,* but of the *theory* from which the hypothesis is derived. On the same page Keuth mentions that in Meehl's own model we might assume that the *theoretical* urn (see below) contains counters designating only "directional hypotheses," but such an assumption would involve the same confusion mentioned above.

Indeed, Keuth is troubled by Meehl's use of an 0.5 probability value and by the model Meehl employs in illustrating its relevance—a model which, Meehl suggests, has certain features in common with actual psychological practice. Now, among other points, Meehl had argued that the point null hypothesis is almost always false in the state of nature ([13], pp. 108–109). Prior to embarking on an examination of the model, Keuth notes that Meehl cannot assign every point hypothesis zero prior probability ([9], p. 539). This is true, but Meehl did not intend to do so; he is careful to state that H_0 is *quasi*-always false, although for practical purposes, "the occurrence of any exact value of X may be regarded as having zero probability" ([6], p. 121).

The model in question consists of two urns. One contains counters designating all actual and possible substantive theories concerning a certain domain of psychology; the other contains counters designating all possible experimental situations involving two groups which are compared on some parameter. Since H_0 is almost always false, each such situation will involve a difference between the two groups with respect to the parameter in question. We then, quite arbitrarily, label one of the groups in each experimental situation the "experimental group" and the other the "control group." Since our labeling is performed randomly, the expected long run frequency should be that half of the time the control group will have a higher *de facto* value for its output variable than will the experimental group, the other half of the time the experimental group's parameter value will be greater than the control group's.

A counter from one urn is then paired *randomly* with a counter from the other urn and the process is repeated indefinitely. Then, by fiat, a theory is declared to be confirmed if the experimental situation is such that the output for the experimental group is greater than the output for the control group, i.e., $\mu_e - \mu_c > 0$. The expected *frequency* of accepting H_1 (H_1 can be stipulated to be either $\mu_e > \mu_c$ or $\mu_c > \mu_e$ across all of the possible pairings) will be $1/2$ and, of course, it would also be $1/2$ for H_2. Before we explore the consequences that should be drawn from this model, let us turn to Keuth's comments.

Keuth begins with the worry that the theoretical urn must contain an infinite

number of counters in order to have counters for each of the values which some random variable might take. Since $n/\infty = 0$ and ∞/∞ is undefined, the probability that a counter representing a *specific* theory or experimental design will be drawn is zero or undefined, depending on the number of counters used to represent a theory. (It seems likely that Meehl intended that a single counter should represent each theory and since nothing hangs on that point, it will be assumed in what follows.) Now if there had to be a different theory for each value that a random variable could take, then indenumerably many counters would be required in the theoretical urn. However, it appears to be Meehl's intention that the actual outcome (e.g. $\mu_1 > \mu_2$) for each experimental situation is all that he wants in the urn containing the experimental situations (and not all the values that the variables could take).

However, for discussion's sake, let us assume that each urn does contain an infinite number of counters. On the basis of his observation concerning the probability of selecting a *specific* counter from either of the urns, Keuth objects that "the probability of subjecting a theory to a relevant test and accepting it is either zero or undefined" ([9], p. 543). Thus, he concludes that this model must be rejected.

But the very point of the model includes arbitrary pairings and capricious stipulations of what is to count as successful theoretical confirmation. Meehl himself calls the urn model "preposterous" ([13], p. 110), and it can be viewed as a *reductio ad absurdum* of certain facets of actual psychological practice. We simply declare, *for purposes of our model*, that a favorable test is one where $\mu_e > \mu_c$; there is nothing wrong with this. Once we do so, the probability of accepting the general hypothesis H_1 is, in the long run, equal to $1/2$, as it is for H_2. Thus, in half the cases we will support our theory. There can be no objections to a model so devised, though, of course, one might object to the uses to which one attempts to put it. Furthermore, since Meehl is concerned with long run frequencies, his model is more apposite than are Keuth's suggested models, one of which involves a two element urn, the other a coin toss ([9], p. 543). We conclude that the urn model is quite coherent. It attempts to show what would happen *if* certain random elements were injected into theory testing; hence these elements are surely not reason for objection.

We may now turn to the "stronger" version of the paradox which Keuth puts by saying that in physics, "if the theory has negligible verisimilitude, the logical probability of surviving such a test [one with great power] is negligible" ([9], p. 539; quoted from [13], p. 113). Rather than discuss this fragmentary statement, let us set out Meehl's full position and Keuth's objections to it. Meehl's position, we believe, is that in psychology, so long as directional hypotheses are employed, perfect power will lead to a probability of at least $1/2$ of confirming our theories *in the long run*.

Let us return to the urn model. There it is clear that in the long run we will have a successful outcome, ($\mu_e > \mu_c$), half of the time. Now, let us suppose, in actual situations there is a deductive link between our theory

and the hypothesis used to test it. In such circumstances the probability in terms of long range frequencies of supporting our theories will depend on the quality of our theories. If we had perfect power and only absolutely true theories, with statistical hypotheses being strictly and deductively derived from them, then we would always end up accepting the hypotheses which flowed from our theories. On the other hand, if our theories always made incorrect predictions, then with perfect power they would never be confirmed. While this is *logically* possible, there is no reason to think that in actuality matters will be more bleak than the chance level of 0.5 described in the urn model. Indeed, Meehl remarks that *even* the psychological theories of his late, uneducated grandmother, possessed more than negligible merit or verisimilitude ([13], p. 111). We take this to mean that even her theories usually made correct predictions more often than not.[3] In actuality, therefore, we might expect that the frequency of accepting statistical hypotheses that tend to confirm our theories will be somewhat greater than 0.5 when perfect power is achieved.

In addition to the problems discussed above, Keuth is worried about Meehl's 0.5 probability value, thinking that it rests on a confusion:

> Possibly Meehl mistook the probability of choosing a hypothesis for test for its probability of being true. This would account for his statement that "if we randomly assign one of the two directional hypotheses H_1 or H_2 [here H_{II} or H_{III}] to each theory, that hypothesis will be correct half of the time" . . . and for his statement "that the effect of increased precision . . . is to yield a probability approaching $1/2$ of corroborating our substantive theory by a significance test, *even if the theory is totally without merit.*" ([9], p. 544)

Keuth is anxious to make the point that, *for any given hypothesis*, the probability of accepting *it* need not be $1/2$. This is certainly correct. However, in the urn model the hypothesis that will be correct half the time is the *generalized* hypothesis that either $\mu_e > \mu_c$ or $\mu_c > \mu_e$, across *all* of the possible urn pairings. A similar sort of point applies to real life situations. Of course the probability of accepting some specific H_1 will depend on the state of nature; any definite hypothesis is either true or false. And, to be sure, to calculate the probability of accepting a given hypothesis we would need to know certain probability values, specifically its prior probability of being true which, in turn, would require that we make what Meehl clearly labels as "illegitimate prior-probability assumptions concerning the actual distribution of true differences in the whole vast world of psychological experimental contexts" ([13], p. 114).

Since Keuth makes so much of this matter, it seems worthwhile to make it clear that Meehl is concerned with frequencies. In a passage *quoted by Keuth* ([9], p. 540) Meehl says:

[3] One could, to be sure, argue this point, but inasmuch as Keuth does not question it, we'll not pursue it here.

> . . . by assuming that there is *no connection whatever between our* theories and experimental designs (the two urn idealization), thereby fixing the *expected frequency* of successful refutations of the directional null hypothesis H_{02} at $p = 1/2$ for experiments of perfect power; it follows that, as the power of our experimental designs and tests is increased . . . we approach $p = 1/2$ as the limit of our *expected frequency* of "successful outcomes." ([13], p. 111; each 'expected frequency' italics ours)

And again, the two urn model

> 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* ([13], pp. 110–111; 'frequency' italics ours).

Other remarks support this reading, and even if Meehl's style was not everywhere so felicitious as it might have been, we see no warrant for alternative readings. Thus we conclude that Keuth is correct in much of what he says in the final three pages of his paper but that it is simply beside the point inasmuch as he is not discussing frequencies (his proposal of a two element urn ([9], p. 543) further suggests this), and Meehl is.

Thus, we conclude that in physics increased power will typically render more difficult the project of confirming a theory, for we can no longer ascribe near misses in its predictions to such matters as measurement error. A point hypothesis is deemed highly improbable when its range is severely constricted by perfect power unless the theory that was used to derive it has more than negligible verisimilitude. In psychology, on the other hand, a point null hypothesis is similar to the point prediction in physics except that instead of wanting to hit the point the psychologist wants to miss it. A directional null hypothesis test does not complicate matters unduly; it simply forces us to conclude that, with perfect power, we will have an expected frequency of at least $1/2$ of accepting the favored hypotheses and thus confirming our theories. Such a high expected frequency cannot add much to our belief that the substantive theory of interest, predicting such an hypothesis, is true; for such hypotheses are reasonably likely to be true and, hence, true whether or not the predicting theory is. This is certainly a difference from the typical case in physics. And although this should be expected once the matter is carefully examined, it is a bit paradoxical inasmuch as it runs counter to our initial intuitions—in the way, say, Skolem's paradox does.

One might conclude from this that in psychology power is to be shunned and, indeed, the authors have heard such a course of action seriously proposed. This would be a drastic measure and, fortunately, there are much sounder ways to circumvent the difficulties. The first step is to heed Meehl's admonition that a directional hypothesis producing a small p value is not "transferrable" to a small probability of "making a theoretical mistake" ([13], p. 107). The meaning of the p value must be kept in proper perspective; it is *at best*

an estimate of the likelihood that a particular statistical hypothesis is true.

Meehl's paper is only one of several recent publications which deals with various problems concerning significance tests. This literature is extensive and the indictments of the *use* (not, of course, the logic) of such tests are severe, so we will not rehearse or argue the problems here. (See the papers in [14], especially [18], [4], [11], and [17].) Though such tests involve numerous putative difficulties, one problem stands out above all others; statistical significance tests as currently used have very little to do with *scientific knowledge* and *explanation.* Mere knowledge of mean differences or correlations pales in comparison with the physicist's procedures. One need not be a rampant Popperian to acknowledge the fact that a fairly likely result cannot go far toward strengthening our confidence in a scientific theory. The problem to which Meehl draws attention at the conclusion of his paper ([13], p. 115), then, is to find ways of testing theories (in psychological research) which go a step beyond current practices, but which are compatible with current capabilities of psychology.

On the prominent view of these tests (as presented above, our examples involve this interpretation) they are not to be viewed as a device which puts an objective probability value on a specific hypothesis to the effect, say, that two means differ,[4] nor are they intended as tools to be employed in modifying our degree of belief in our hypothesis. Rather, they are simply decision procedures, advising when we may accept (not automatically believe as true) a specific hypothesis, given that we have previously decided that we are willing to chance committing a Type I error so many times out of a hundred (this is the view associated with the names Neyman, Pearson and Wald). On this view a p value is indeed a probability value, but it represents the *relative frequency* with which we would obtain results like the current one (say, rejection of a directional null), if, *per impossible,* we performed many literal replications of our experiment (on types of replication, see [11]).

It seems to us that such a *decision* to accept or reject a hypothesis can have very little to do with scientific knowledge;[5] it may be useful in making certain decisions, but it does not seem to be the answer when we are concerned with what we should rationally believe.[6]

But if we could assign, even quite roughly, some probability value to our statistical hypothesis (which most experimenters tend to do implicitly anyway) it seems easier to imagine its relevance. To put it baldly, if we are interested

[4] Most researchers ignore this point and this helps generate the problem Meehl discusses, for, of course, if we were only engaged in decision making we might entirely avoid problems concerning confirmation. But, like Meehl, we are concerned with actual practices in psychological research.

[5] Tversky and Kahneman [21] have indicated that a reliance on a strict decision procedure can result in an undesirable situation in which two separate tests of a certain statistical hypothesis result in contradictory conclusions, but if the data were combined into a single test, the hypothesis would receive stronger and unqualified support.

[6] *At the very least* one would need, in addition to the tests, some sort of concept of "epistemic utilities," though we know of no wholly adequate and cogent theory of these; see [7] and [10].

in whether or not we should accept or strongly believe[7] in a substantive theory, we need at least a vague idea about the probabilities of the relevant evidence.

Just how one should make such assignments, even in a loose way, is a very complicated problem involving fundamental issues in statistics and probability theory. One could do so by adopting a Bayesian approach to statistics; such an approach generally involves a subjective or personalistic interpretation of probability (for discussions see [5], [20]). This approach, though we believe it to be quite promising and exciting, is not without problems of its own; moreover, many will find its use of subjective probabilities suspect. Alternatively, one might acquiesce in assigning something like probabilities to single cases, using what Reichenbach calls "posits," ([16], pp. 313 ff.; see also [19], pp. 93 ff.) or, here one might countenance handling the single case by utilizing something like a Popperian "propensity" interpretation of the probability calculus. However, we are not interested so much in how one obtains these values, though of course this is an important and fundamental question; rather we are concerned to make the point that we need some way, rough though it may be, of knowing how likely it is that our statistical hypothesis is true.

Once we get this far, what are we to *do* with the probabilities that our statistical hypotheses are true? Our primary concern, surely, is to make them bear on *substantive* theories. Now although many objectivists, and even some subjectivists—including De Finetti himself—have refused to assign probability values to theories,[8] it seems to us that, at the very least, approaching matters as if we wanted to make such assignments can be a very helpful way of dealing with the problem at hand. At a minimum, taking such an approach and utilizing Bayes' theorem possesses a definite heuristic value in our search for ways to subject our theories to more fruitful tests.

Assuming discreteness, a simple form of Bayes' theorem can be written thus:

$$Pr(T/E) = \frac{Pr(T)Pr(E/T)}{Pr(E)}$$

That is, the posterior probability of a theory—$Pr(T/E)$—equals the prior probability of the theory multiplied by the likelihood of the evidence on the theory, all divided by the prior probability of the evidence. Thus, if we take the prior probability of T to reflect our degree of belief in the

[7]We do not here propose to enter the sticky debate over whether or not it is the role of the scientific enterprise to produce knowledge or merely to change our degrees of belief in certain theories, that is, over the problem of rules of acceptance in inductive logic. Since nearly all positions see a high probability value as a necessary, though not sufficient, condition for performing detachment, we may here restrict ourselves to the problem of assigning probabilities to theories.

[8]Many subjectivists, notably [5] have disagreed; more surprisingly, perhaps, so have certain objectivists, see [19], [12].

theory, the theorem, together with values for the other expressions, shows us what our degree of belief should be in T, given some bit of evidence.[9]

This approach to theory confirmation does not depend on any specific interpretation of the probability calculus and it has been defended by objectivists ([19], [12]) and subjectivists [8] alike, although, of course, implementation of the method, e.g., determining values for our prior probabilities (the so-called "priors"), may vary depending on which interpretation one endorses. In addition to this neutrality, Bayes' theorem has advantages over other traditional approaches. For example, the hypothetico-deductive model assumes that we *deduce* our predictions from our theory, that is, that the likelihood of the evidence on the theory be 1.0. This is often a rather unrealistic picture, especially in the social sciences, and it is a picture which Bayes' Theorem does not force upon us (cf. [12] for a good discussion of this). The theorem also differentiates between weak and strong pieces of evidence; moreover, it reflects the fact that some theories are from the start inherently more plausible than the others. As we and many others have noted, if a certain piece of evidence would be expected come what may, it can hardly give much support to any specific theory. On the other hand, if the evidence would be unlikely *unless* the theory predicting it is true, it is a strong piece of support for our theory. We will be more impressed by a theory which correctly predicts that children in suburbia will have a mean I.Q. of between 101 and 103 than by a theory which successfully predicts that their I.Q. will differ (though by no *specified* amount) from the I.Q. of ghetto children. A cursory examination of Bayes' theorem shows that it captures this insight. Finally, we do not require precise probability values for our theorem; as with most schemata of this sort, topological considerations suffice. (For a vigorous defense of the usefulness of Bayes' theorem in investigating theory confirmation see [19] or [12].)

The theorem follows intuition and the counsel of many previous philosophers of science in suggesting that, whenever possible, we should make tests of a theory which involve predictions which would be surprising if the theory were false. In this context we can profitably look back at Meehl's paradox. As Meehl notes ([13], pp. 112–113), the paradox arises because physics and psychology currently make very different sorts of predictions. The prior probability of a typical psychological theory's predictions panning out are reasonable high while those of a physical theory's predictions being correct are, by comparison, reasonably low—the reason being that the prior probability of the evidence varies in the two cases. Put another way, a successful point prediction gives us much more information; by ruling out numerous possible alternative states of affairs it leaves us with a better idea of what the real state of affairs is and of which theory is true.

[9]Matters, of course, are complicated by the fact that in general the evidence can be said to obtain with a probability less than 1.0. In a full treatment, this would have to be faced and, e.g. a rule of acceptance for hypotheses proposed or more complicated mathematics introduced.

We now turn to procedures which, if more widely adopted, would help raise theory confirmation in psychology above the primitive state in which the widespread use of only directional hypothesis testing now leaves it.

One strategy, not infrequently offered, counsels us to employ confidence levels. The reason for so doing, though less frequently appreciated, is this: they approximate point predictions. A theory that predicts that a certain parameter will fall within a certain confidence interval in effect rules out a good many possible states of affairs and by so doing rules out a good many competing theories, though certainly not so many as does a point prediction. Put another way, *ceteris paribus*, the prior probability of the evidence is generally much smaller here than it is in the case of a simple directional difference between means. So, hopefully, psychologists can devise theories which, though incapable of making point predictions, can generate hypotheses more specific than those which state merely that a specified directional mean difference exists. Here confidence intervals are an obvious half-way house. The more psychology progresses and the better its theories get, the smaller we may hope the predicted intervals will become.

Sometimes it is suggested that confidence intervals are important because they help us locate large differences between means, differences of sufficient magnitude to be of "theoretical or practical importance" ([22], p. 203). There is indeed something to this, but one must be careful not to value large differences for the wrong reason—such differences are not intrinsically more *theoretically* important than small mean differences. And a theory which predicted a very small mean difference but did so accurately could not be thereby faulted. But there *is* a reason why large mean differences can provide more theoretical confirmation than small mean differences.

If we find a large difference, it seems to us that there are far fewer *relevant* alternative theories that could conceivably account for it than if our difference were small. Technically there are many theories that could account for differences of either size, but as a practical matter we would tend to assign much lower priors to many of those that predict a large mean difference (this would be especially easy for the subjectivist). An example should help clarify the point.

Suppose we have an experimental situation in which one group of subjects is given a new stimulant drug and another is given a placebo. Since we cannot counterbalance everything (time, location, experimenter, etc.), the experimental conditions will be somewhat different for the two groups.[10] All these small and seemingly irrelevant differences can conceivably make some small difference in the amount of anxiety either our experimental group or our control group experiences. (Perhaps we are forced to use different investigators and one wears a bright chartreuse tie—it could plausibly be thought to provoke a little anxiety, but not a lot.) However, they could probably not be thought to cause a large difference, whereas the drug might. (Of

[10] Clearly the problem is amplified considerably in nonexperimental settings.

course, there are small differences in nature, and a *complete* theory will have to predict them as well as large ones.)

Another useful strategy is suggested by Lykken [11], in his discussion of replications. As he points out, chances of successfully replicating an experiment several times are not as good as the chances of one successful run. Replication is always important, though replications of a phenomenon involving simply a putative directional mean difference cannot go a long way in helping us narrow in on the actual state of affairs and so cannot rule out a lot of competing theories.

Another strategy is this. Psychologists can, we think, realistically hope to devise theories which make predictions about range or ordinal positions of several numerical values. Here again, such predictions are not apt to be successful come what may. They will not be so likely to pan out unless our theory which makes the predictions is reasonably close to being true. If our theory makes the prediction that four different groups should have mean scores which fall into a specified order, the prior probability of such an outcome is reasonably low compared to that of mere difference in the means, and if the groups do rank as predicted there is reason to put some faith in our theory. Of course we cannot have a specific prior probability that applies to each case; as always, the quality of the evidence can vary. A prediction, that twenty-year-olds can run faster than forty-year-olds who can run faster than sixty-year-olds, would not be of great interest. Our point, though, is general and comparative. *Ceteris paribus,* such a strategy is always preferable to use of the directional null hypothesis.

Yet another strategy is this. Psychologists might profitably attempt to formulate theories which make predictions about the forms of certain curves, that is, theories which predict that a certain relationship between variables could be captured by a function of a certain sort. For example, a theory might predict that a linear function would represent a certain relationship, though the theory need not be so precise as to predict a slope constant or the value of an intercept.

All these suggestions, of course, are but variations on a single theme: attempt to make predictions which are precise enough to rule out competing theories and, thus, which are capable of confirming the theory of interest to some degree. Clearly, such strategies will only be useful within the context of a large, ardous, full-blown research program involving large scale cooperation—scattered and isolated bits of research can never do much to support a theory.

These strategies are more than attractive but unattainable goals. Much of psychology is currently at a point where it can reasonably attempt to make the sorts of predictions we have mentioned, though there can be no cookbook for devising either the theories which can make some predictions nor the predictions themselves. For this, creativity and, perhaps, even genius is required. One thing, however, is certain; unless psychologists see the value of devising such theories, they will not be likely to try. Perhaps one way

to achieve such goals is suggested when we look to physics where we see how a few laws and concepts tie together so many and varied phenomena; we are then bound to be struck by the fact that a few underlying theoretical entities account in large measure for its unity, integration, and predictive and explanatory power. It is reasonable to suspect that a wider use of theoretical concepts would prove useful in psychology as well, but only the actual devising and testing of theories will tell for certain.

REFERENCES

[1] Bakan, D. "The Test of Significance in Psychological Research." In *On Method.* San Francisco: Jossey Bass, 1967. Pages 1-29. (Reprinted in [14])
[2] Bandt, C. L. and Boen, J. R. "A Prevalent Misconception about Sample Size, Statistical Significance, and Clinical Importance." *Journal of Periodontology* 43 (1972): 181-183.
[3] Bolles, R. C. "The Difference between Statistical Hypotheses and Scientific Hypotheses." *Psychological Reports* 11 (1962): 639-645.
[4] Camilleri, S. F. "Theory, Probability, and Induction in Social Research." *American Sociological Review* 27 (1962): 170-178. (Reprinted in [14])
[5] Edwards, W., Lindman, H. and Savage, L. J., "Bayesian Statistical Inference for Psychological Research." *Psychological Review* 70 (1963): 193-242.
[6] Hays, W. L. *Statistics for the Social Sciences.* (2nd ed.). New York: Holt, Rinehart and Winston, 1973.
[7] Hempel, C. G. "Deductive-Nomological vs. Statistical Explanation." In *Minnesota Studies in the Philosophy of Science.* Vol. III. Edited by H. Feigl and G. Maxwell. Minneapolis: University of Minnesota Press, 1966. Pages 98-169.
[8] Jeffreys, R. *The Logic of Decision.* New York: McGraw-Hill, 1965.
[9] Keuth, H. "On Prior Probabilities of Rejecting Statistical Hypotheses." *Philosophy of Science* 40 (1973): 538-546.
[10] Levi, I. *Gambling with Truth.* Cambridge, Massachusetts: M.I.T. Press, 1967.
[11] Lykken, D. "Statistical Significance in Psychological Research." *Psychological Bulletin* 70 (1968): 151-159. (Reprinted in [14])
[12] Maxwell, G. "Corroboration with Demarcation." In *The Philosophy of Karl Popper.* Edited by P. A. Schilpp. LaSalle, Illinois: Open Court, 1974. Pages 292-321.
[13] Meehl, P. E. "Theory Testing in Psychology and Physics: A Methodological Paradox." *Philosophy of Science* 34 (1967): 103-115. (Reprinted in [14])
[14] Morrison, D., and Henkel, R. E. (eds.). *The Significance Test Controversy: A Reader.* Chicago: Aldine Publishing Co., 1970.
[15] Popper, K. *Conjectures and Refutations.* London: Routledge and Kegan Paul, 1963.
[16] Reichenbach, H. *Experience and Prediction.* Chicago: University Press, 1938.
[17] Rosenthal, R. and Gaito, J. "The Interpretation of Levels of Significance by Psychological Researchers." *Journal of Psychology* 55 (1963): 33-38.
[18] Rozeboom, W. W. "The Fallacy of the Null Hypothesis Significance Test." *Psychological Bulletin* 57 (1960): 416-428. (Reprinted in [14])
[19] Salmon, W. *The Foundations of Scientific Inference.* Pittsburgh: University Press, 1966.
[20] Savage, L. J., *et al. The Foundations of Statistical Inference.* New York: John Riley and Sons, 1962.
[21] Tversky, A. and Kahneman, D. "Belief in the Law of Small Numbers." *Psychological Bulletin* 76 (1971): 105-110.
[22] Woodson, C. E., "Parameter Estimation vs. Hypothesis Testing." *Philosophy of Science* 36 (1969): 203-204.