

# Robustness and Integrative Survival in Significance Testing: The World's Contribution to Rationality\*

J. D. TROUT

---

## ABSTRACT

Significance testing is the primary method for establishing causal relationships in psychology. Meehl [1978, 1990a, 1990b] and Faust [1984] argue that significance tests and their interpretation are subject to *actuarial* and *psychological* biases, making continued adherence to these practices irrational, and even partially responsible for the slow progress of the 'soft' areas of psychology. I contend that familiar standards of testing and literature review, along with recently developed meta-analytic techniques, are able to correct the proposed actuarial and psychological biases. In particular, psychologists embrace a *principle of robustness* which states that real psychological effects are (1) reproducible by similar methods, (2) detectable by diverse means, and (3) able to survive theoretical integration. By contrast, spurious significant findings perish under the strain of persistent tests of their robustness. The resulting vindication of significance testing confers on the world a role in determining the rationality of a method, and also affords us an explanation for the fast progress of 'hard' areas of psychology.

### 1 Introduction

### 2 The Charge that Significance Tests in Psychology are Unintelligible

### 3 The Development of Corrective Meta-Analytic Techniques

### 4 The Robustness of Psychological Phenomena

### 5 Submission and Editorial Biases

### 6 Conclusion

---

## I INTRODUCTION

Cognitive research on the frailties of human judgment (for a recent review, see Arkes and Hammond [1986]) has quickly made its way into the philosophical

\* I would like to thank Dick Boyd and Phil Gasper for helpful comments on the ideas presented here.

literature on rationality (Cherniak [1986]; Davidson [1976]; Giere [1988]; Goldman [1986]; Stich [1990]). If you subscribe to the new naturalism in epistemology, you may not like what you see, at least initially. Lay people and clinical specialists commit systematic errors when reasoning inductively.<sup>1</sup> These charges are now casting a pall over the more theoretical disciplines as well, particularly psychology and sociology, where research draws heavily on tests of statistical significance. It is alleged in some quarters that significance testing is so problematic 'the usual research literature review is well-nigh uninterpretable' (Meehl [1990b], p. 197), a conclusion that 'is precisely the right response of a rational mind' (p. 198) and one which constitutes 'a methodological problem of our field that I insist is not minor but of grave import' (Meehl [1990a], p. 108). (For similar charges elsewhere in Meehl's work, see [1967], [1978], and [1985].)

With the rationality of this central scientific practice now under attack, the principles and purposes of significance testing deserve closer examination in the philosophical literature than they have so far received. I don't propose to address all of the reasons that have been given for suspicion concerning significance tests (Morrison and Henkel [1970] provides a very thorough treatment of the issues). Although the methodology of significance testing is often blamed for the slow progress of 'soft' (social, personality, counseling, clinical, and educational) psychology, the criticisms of significance testing typically are not presented as depending on the subject matter. Rather, the charge is posed in quite general and neutral terms; significance testing is 'a poor way of doing science' (Meehl [1978], p. 806).<sup>2</sup> Following a survey of the charges, I will mobilize three reasons for thinking that the critic's concerns are either peripheral or misleading.

First, to the extent that significance tests could be applied with clearer principle, and interpreted with greater rigor, there are meta-analytic techniques, much discussed in the theoretical literature, designed to refine significance testing in just these ways. Second, the criticisms examined here ignore the methodological and editorial assumption that real psychological phenomena are *robust*, and thus their effects are (1) reproducible by maximally similar methods, (2) detectable by diverse instruments and measurable by diverse scales, and (3) able to survive attempts at theoretical integration. By contrast, spurious significant findings—effects which reach conventional standards of significance by chance—don't survive in the literature, because

<sup>1</sup> The discussion that follows will focus entirely on errors in inductive reasoning. For experimental evidence from the psychology of human inference indicating the systematic violation of deductive canons as well, see Cherniak [1986] and Wason and Johnson-Laird [1972].

<sup>2</sup> In Meehl's work, for example, the formal, statistical issues are discussed independently of such substantial issues as the imprecise predictions of 'weak' theories. I will examine the statistical problems, and Meehl [1978] covers these in a motley of up to 20 formal and substantial problems in the so-called 'soft' areas of psychology.

they don't survive in the lab or in the field. Finally, the fact that various factors of unknown value influence measurements is not a problem peculiar to significance testing; rather, it is an expression of the more general fact that we fall short of omniscience, a condition that hasn't prevented scientific progress elsewhere.

Continued reliance on significance testing, then, should not be deemed epistemically irresponsible. Far from having established the psychologist's irrationality, the critic of significance testing bears an additional burden; the critic must explain how the 'harder' areas of perceptual and cognitive psychology, indebted as they are to significance testing, could have enjoyed such noticeable progress in the last 50 years or so if the methodology of significance testing were as defective as the critics allege.

## 2 THE CHARGE THAT SIGNIFICANCE TESTS IN PSYCHOLOGY ARE UNINTELLIGIBLE

It is primarily in terms of statistical significance tests that research results are reported in fields such as psychology and sociology. Those most commonly used are t tests, F tests, and analyses of variance and of covariance. Let's look at one standard statistical strategy. A *population* of subjects is identified, and one large group, drawn from that population, is selected and divided into two subgroups or *samples*. Each sample is given a different experimental *treatment*, and subject performance under each treatment yields respective sample means. What we want to determine is whether differences between subgroup means are large enough to be attributed primarily to the influence of experimental conditions rather than to chance factors deriving from the assignment of subjects into the subgroups. Variation within the subgroups is used to estimate the amount of variation attributable to chance factors, such as sampling error. With the sampling error now estimated, we can determine whether the difference in subgroup means due to experimental factors is great enough to merit treating the two subgroups, originally drawn from a common population, as representative of two different populations.

The hypothesis tested is that the experimental treatment *did not* have an influence on subjects sufficient to support the conclusion that a significant difference exists between the two samples. When the treatment is understood as an independent variable and the influence or effect as a dependent variable, this hypothesis amounts to the projection that no relationship exists between two or more variables being examined. This is the *null hypothesis* ( $H_0$ ). Strictly speaking, a significance test attempts to identify the *risk* associated with the rejection of the null hypothesis. In light of the particular standard error, the significance test estimates the probability that the difference in subgroup means would be repeated if two subgroups were drawn again from that population. The standard level of risk or significance is normally set at 0.05;

that is, the null hypothesis can safely be rejected if the risk of duplication due to chance falls below 0.05. The rejection of the null hypothesis is normally interpreted as confirming the alternative hypothesis ( $H_1$ ).

In relying on these particular significance tests, we risk two sorts of error. We commit a Type I error if we reject the null hypothesis when it is true. In such a case, the effect has reached significance by chance variation in sampling and is not, as we sometimes say, 'real'. A Type II error occurs if we fail to reject the null hypothesis when it is false. Here, again due to chance variation in sampling, no effect materializes even though the two variables are in fact related.

I will be concerned with two factors which complicate the interpretation of significance tests. One factor is largely *actuarial*, and the other, *psychological*. Both incline the investigator to Type I errors. The first complicating factor derives from the simple repetition of experiments on the same dependent variable.<sup>3</sup> An effect which reaches the standard 0.05 level of significance (taken as indicating that there is a 95 percent chance that the effect is indeed real) is reported as significant. Therefore, repeated experiments on the same dependent variable are likely to produce Type I errors, generating results which are spuriously identified as significant, even where there is no real relationship between the independent and dependent variables.

The second complicating factor trades on the first.<sup>4</sup> There is body of psychological research suggesting that psychologists (and other scientists) are subject to an 'availability bias', causing them to underestimate the frequency with which 'spurious significant' findings get published. Faust [1984] draws on recent research in cognitive psychology which attempts to demonstrate such systematic errors in both lay and scientific reasoning. The 'base-rate problem' is one of the most prominent of these results. People normally underutilize important information about the relative frequency at which a certain event, object, property, *etc.*, occurs in a population.

In a ground-breaking study, Kahneman and Tversky [1973] constructed personality descriptions of five people putatively sampled from a class of 100 professionals—engineers and lawyers. For each description, subjects were asked to estimate the probability that it belonged to an engineer rather than a

<sup>3</sup> Talk of 'replication' or 'repetition' here must be qualified. A replication is an instantiation of the same design, from a theoretical view. There are a variety of differences between the original experiment and the 'replication' that are unavoidable, even when scientists explicitly attempt to replicate an experiment. Original experiments and their replications are normally carried out at different times, in different settings, often using different technicians. With the qualification now registered that perfect replication is impossible, we can at the same time acknowledge that this limitation seldom presents an insuperable problem. The particular theory tells us which differences *make* a difference *e.g.*, is likely to produce a unique artifact, *etc.*)

<sup>4</sup> As a matter of fact, Meehl's groundbreaking work [1954] on the prediction of performance using actuarial rather than clinical (*e.g.*, configural, stereotype, *etc.*) guidelines anticipated much of the recent work on the importance of base-rate information.

lawyer. Subjects were told in one condition that 70 percent of the persons described were lawyers, and 30 percent were engineers. In the second condition, the frequencies were reversed. Despite the fact that these personality descriptions were not particularly diagnostic of profession and that under such conditions base-rate information should have had a powerful influence on the probability estimates, subjects ignored base-rate information; instead, they appeared to rely on the stereotype (personality) description, violating actuarial rules in predicting the relevant occupation.

It is not as though subjects were incapable of appropriately exploiting base-rate information. In the absence of any personality information, for instance, subjects used prior probabilities correctly, estimating that an individual randomly selected from the 70 percent lawyer sample had a 0.7 likelihood of being a lawyer, and so on for the other assignments. However, even when the following, *entirely nondiagnostic* personality information was presented, subjects once again ignored the base rate:

Dick is a 30-year-old-man. He is married with no children. A man of high ability and high motivation, he promises to be quite successful in his field. He is well liked by his colleagues (Kahneman and Tversky [1973], p. 242).

Here, subjects rated the engineer and lawyer likelihoods to be equal—in other words, 50 percent—whether the subjects were provided the 70 percent base rate or the 30 percent base rate. So it appears that, as a class, people respond differently to the *absence* of information than they do to *worthless* information; when *no* specific evidence is provided, prior probabilities are correctly appreciated, and when useless information is provided, prior probabilities are ignored. The moral normally drawn is that people must suffer a woeful lack of appreciation of the importance of base-rate information if they can be so easily distracted by such worthless ‘stereotype’ information.

L. Jonathan Cohen [1986] was the first to raise serious and systematic doubts about this pessimistic interpretation of the experimental results. According to Cohen, the responses of statistically untutored people reveal that they can be led astray in their application of inductive principles; if we recognize a distinction between counterfactualizable and noncounterfactualizable conceptions of probability, and the ordinary dominance of the former over the latter, experimental subjects’ violation of normative rules can be rationalized. Cohen concludes that these experiments therefore do not demonstrate human irrationality.

Nothing in the present paper, however, depends on the truth of Cohen’s conclusion. Even if, contrary to Cohen’s contentions, lay persons do not commit systematic errors of inductive inference, this (putative) psychological bias is subject to correction by the forces described in the sections that follow.

Nor can we secure the integrity of inductive scientific inference by locating the defect entirely within lay reasoning. Faust ([1984]; also see Chapman and

Chapman [1969]) argues that the underutilization of base-rate information is not peculiar to lay judgment; insensitivity to prior probabilities produces systemic and fundamental errors in routine *scientific* judgments as well.<sup>5</sup> As a result, methodological practices such as the literature review are subject to the charge of irrationality if they depend on significance testing, and fall in with the evaluative standards for publication that currently prevail.

The way to the base-rate fallacy in psychology is abetted by two institutional biases, one in *article submission* and another in *editorial acceptance*. Researchers typically do not submit articles reporting null results, and journals typically reject such articles in the relatively rare instances in which they are submitted. As a result, the journal audience is presented with a population of studies badly skewed toward those reaching significance. And to make matters worse, the audience possesses no base-rate information concerning the frequency with which experimental efforts *failed* to reach significance; such efforts get filed away, and are thus unavailable to the journal audience.<sup>6</sup>

It is for these reasons that David Faust cautions those who rely on significance tests in psychology, and attributes their blind persistence to an ‘availability bias’:

If base rates are ignored, one might assume that there is a 95 percent chance that the results are indeed significant. If base rates had been considered, it might be shown that the finding of significant results, regardless of external reality, was almost ensured. . . . The far greater availability of concrete instances of significant findings further works against the formation of accurate impressions [of base rates]. . . . That so little attention has been directed towards gathering base rates for the occurrence of nonsignificant results may also indicate a general failure to appreciate the crucial importance of base rate information (Faust [1984], pp. 93–4).

It has thus grown popular to accuse psychologists with, at best, epistemic irresponsibility and, at worst, epistemic bad faith. Faust seems inclined towards the charge of epistemic irresponsibility, while Paul Meehl is less generous. The most persistent figure behind this accusation, Meehl speculates that students and colleagues haven’t responded to his warning ‘since they might not know what to do next if they took it seriously’ ([1990b], p. 198).

<sup>5</sup> Lugg [1987] objects to Faust’s move from evidence of error in clinical judgment to error in scientific judgement. Although Lugg is correct to remind us of the presumptive and impressionistic character of Faust’s argument, the sorts of *practices* targeted by Faust, such as literature review, are not peculiar to clinical fields. Nor does the hard science/soft science distinction appear to mark a relevant difference in the standards of literature review (Hedges [1987]).

<sup>6</sup> On top of it all, pilot studies—which are, after all, experiments in the small—are never counted among the total population of studies, further increasing the total number of experiments executed, and thus the likelihood of producing a spurious significant finding.

And just in case his diagnosis of denial has not been digested, Meehl adds the following:

As is well known, were some mischievous statistician or philosopher to say 'Well, five times in a hundred you would get this much of a difference even if the rats had learned nothing, so why shouldn't it happen to you?', the only answer is 'It could happen to me, but I am not going to assume that I am one of the unlucky scientists in 20 to whom such a thing happens. I am aware, however, that over my research career, if I should perform a thousand significance tests on various kinds of data, and nothing that I researched had any validity to it, I would be able to write around 50 publishable papers based upon that false positive rate' (Meehl [1990b], p. 212).

Now, the problems Meehl describes here are real, but they are not as serious (nor as wilfully committed) as Meehl thinks. Meehl's concerns no doubt derive from his statistical sophistication, along with frustrated efforts to improve the testing of 'weak' (typically data-driven) theories by largely methodological means. Unfortunately, it is easy to get the impression from the above passage that researchers who rely on significance testing are either dishonest or stupid. The honest scientist is left with little recourse but to abandon or radically reform significance testing, opting for other measures such as curve-fitting (or perhaps adopting stricter standards of significance), while the rest must remain in the grip of a suspect methodology.<sup>7</sup>

The problem here, of course, is that we can't tell in advance whether a particular effect reported as significant is authentic or spurious. The value of these actuarial and (alleged) psychological distortions are unknown at any given time, and this is seen by critics as a serious problem, rather than just as marking the boundary of our own knowledge at the time. According to Meehl, this uncertainty can be reduced by adopting more stringent testing and review standards. Investigators should predict not just an effect, but an effect *size*, set

<sup>7</sup> This impression is aided by the critics' distorting tendency to ignore important aspects of article submission and evaluation, as well as the actual availability of null results. These omissions allow the criticism of significance testing to be presented in the most dramatic and threatening light. But this impression is inaccurate. Articles reporting a single experiment are relatively uncommon. Normally, an article reports two or more experiments, so it is doubtful that 50 experiments that reach significance would yield an equal number of journal articles. Numbers are important here because, first, the psychological plausibility of a hypothesis is sometimes presented as a partial function of the sheer number of articles supporting it, and second, the psychologist's complicity in ignoring the obvious threat of spurious significance, the critic insinuates, should be explained in terms of the psychologists' desire for a lengthy list of publications. Nor are concrete instances of null results unavailable, and thus, as Faust implies, involved in 'desensitizing' the psychologist to the frequency with which null results occur. Researchers are reminded of null results every time their experiments support the null hypothesis, every time such results are described in personal communication, and every time they are reported among a series of published experiments. Finally, null results sometimes derive from poor design, and referees often reject articles on the basis of poor design. The critics discussed here, however, present the referee's behavior in such cases as a simple reflection of a bias against accepting studies reporting null results, rather than as a reliable disposition to reject poorly designed studies.

their statistical power at 0.9, and report all pilot studies and attempts at replication. Journal editors should generally require a successful replication, tables should report means and standard deviations, and a section of the journal should be devoted to the publication of negative pilot studies. Literature reviewers should, wherever possible, mention the statistical power of the included studies, and always avoid the reporting of 'box scores', tabulating the number of studies on a particular issue reaching significance and those failing to reach significance, in an effort to estimate a theory's verisimilitude.<sup>8</sup> Finally, theoreticians should attempt to develop alternative methods of deriving predictions from weak theories.

Many of these recommendations would indeed allow us to avoid Type I errors, but such standards have a cost that the above critics have omitted from the discussion. Presumably we could always further reduce the risk of Type I error by demanding an even greater number of replications, count *all* pilot studies among the total run, and literature reviewers could always avoid being drawn in by Type I errors by assuming a rather large impact for an alleged bias in submission and editorial acceptance. But by doing so, we risk committing more Type II errors—due to designs and reviewing standards which, once reformed, would be insensitive to many real causal factors.

It is worth noting that this actuarial criticism is presented as domain-neutral; it is the *statistics* of repeated experimentation (on the same variable), not the subject matter, that guarantees the production of this study artifact of actuarial bias. Curious, then, that the critics should selectively attack only a few areas of psychology (*e.g.*, social, personality, counselling, clinical, and educational) rather than raising quite general doubts about the integrity of all

<sup>8</sup> This practice of tabulating box scores is mistaken, scolds Meehl, because it assumes that a significant result does a theory as much good as a null result does it damage. The falsity of this assumption is exposed, Meehl continues, in a moment's reflection on the dynamics of theory testing: "This is scientifically a preposterous way to reason. It completely neglects the crucial asymmetry between confirmation, which involves an inference in the formally invalid third figure of the implicative syllogism (this is why inductive inferences are ampliative and dangerous and why we can be objectively wrong even though we proceed correctly), and refutation, which is in the valid fourth figure, and which gives the *modus tollens* its privileged position in inductive inference" ([1978] p. 822; also see [1990b, p. 233]). It's not clear whether Meehl is simply making the point that, when the hypothesis is precise and a test of it is well designed, null results weigh more heavily against the theory than positive results support it, or whether he also intends the appeal to the favored status of *modus tollens* in theory testing and the blithe use of 'corroboration' and 'refutation' as an endorsement of falsificationist philosophy of science. Since many of those party to the significance testing dispute are psychologists unfamiliar with the recent developments in the philosophy of science, it might be worth mentioning in such contexts that the last generation in philosophy has witnessed the erosion of the falsificationist program, and that there are few real adherents left. The interested scientist might begin with Putnam's now classic "The 'Corroboration' of Theories", in Putnam [1975].

disciplines—psychological and nonpsychological alike—that employ tests of statistical significance. For instance, this same tendency toward Type I error pervades the significance testing methodology of the quickly advancing fields of perceptual and cognitive psychology, yet the critic doesn't impugn those mature and successful domains. As I will argue later, scientists typically either possess knowledge of corrective techniques, are subject to institutional standards of improvement applied at the journal referee stage, or change their views in light of feedback from the world; they don't remain forever innocent of nature's verdict. But for present purposes, the critic is simply mistaken in suggesting that psychologists are unaware of the threat of Type I error attendant upon the kind of null-hypothesis testing currently employed. In fact, a separate body of research has arisen in order to supply the tools to correct the tendency toward Type I error due to actuarial bias.

### 3 THE DEVELOPMENT OF CORRECTIVE META-ANALYTIC TECHNIQUES

The primary source of statistical vigilance here is meta-analysis, a field that has grown in response to the increasing recognition that literature summaries are initially subject to the actuarial and psychological biases detailed above. Meta-analysis is the study of sets of research results, normally carried out on a number of studies regarding the same research question.

As we saw, psychology journals represent a biased sample of the total number of studies performed on a particular issue. By sheer repetition of studies (on the worst scenario), 5 percent of the studies reported in psychology journals represent Type I errors, and 95 percent of the studies showing nonsignificant results (set at, *e.g.*,  $p > 0.05$ ), are filed away in lab drawers, not regarded as worth reporting or even submitting for publication. This state of affairs, presented by Meehl and Faust as an *unrecognized or unacknowledged* source of irrational judgment, is so well known among researchers that it even has a name: The File Drawer Problem (Rosenthal [1979]). In order to protect against committing a Type I error due to simple repetition, we must not underestimate the number of studies carried out on a particular question.

Meta-analysis is particularly well suited to correct such sampling biases predictable from simple actuarial rules. As a remedy for the file drawer problem, meta-analysis offers the Tolerance Table. The Tolerance Table tells us how many new, filed, or unretrieved studies with a null mean it would take to drag down a population of significant studies (of a certain mean and size) to just below the conventional standard of significance (See Rosenthal and Rosnow [1984], pp. 378–82). For any substantial population of studies

reaching significance, the number of null studies it can tolerate is (typically) surprisingly high.<sup>9</sup>

Meta-analysis addresses a variety of other potential artifactual distortions, deriving from sampling error, treatment effects, inadequate statistical power, or the strict falsity of the null hypothesis when interpreted literally (known as the 'crud factor').<sup>10</sup>

#### 4 THE ROBUSTNESS OF PSYCHOLOGICAL PHENOMENA

Given the promise of meta-analysis, we now have reason to think that the biases predictable from actuarial sampling data can be corrected statistically. Although this assurance is encouraging, success does not depend on it. There are properties distinctive of effects that are real rather than spurious. In the process of theory construction and literature review, scientists assess the plausibility of theories, and the reality of postulated entities, on the basis of the robustness and integrative survival of reported effects.

Nearly all scientists assume that real psychological phenomena are *robust*; they emerge and persist under a variety of independent measurements or methods of detection applied to them. It's not surprising, then, to discover scientists employing a certain *robustness principle* in their literature reviews: Effects that reach significance are more likely to be real if they have been arrived at by diverse, independent methods, and reproduced by maximally similar methods. The assumption of robustness is evident in attempts to 'triangulate' on research results. If only one procedure is used, the resulting value could always be an artifact of one type or another, and we have no way of checking this possibility. However, if two or more operational procedures are used, the different peripheral causal influences attendant on each could be thought to produce divergent results—were those different techniques not measuring the same persistent object. In order to confirm the existence of this object, its presence must be established independently and variously.<sup>11</sup>

<sup>9</sup> Rosenthal [1979] reports that, of 94 experiments on interpersonal self-fulfilling prophecies summarized in Rosenthal [1969], it would take 3,263 unreported studies averaging null results (where the average mean, measured in standard deviation units, equals 0.00) before we are forced to conclude that the overall outcomes are the result of a sampling bias of the studies summarized in the review. A more recent review (Rosenthal [1976]) of 311 studies of the same area raises the stakes considerably: 49,457 unreported studies are tolerable. As either the population of available studies increases, or the mean ( $Z$ ) score (in the same direction) increases, the hypothesis that full file drawers could explain such a mean is itself implausible.

<sup>10</sup> The crud factor is especially influential in the case of large samples. As sample size increases, so does the probability that correlations among the variables under investigation will reach the conventional level of significance (see Lykken [1968]; Meehl [1967, 1990a,b]). Although the crud factor does not represent a Type I error—these variables are really correlated—it is still a prominent study artifact. Hunter and Schmidt [1990] proposes corrective measures for this artifact and the others mentioned above; also, see Green and Hall [1984].

In the history of chemistry and physics, attempts to establish the existence and nature of molecules clearly triangulated on diverse research areas and techniques. Dissimilar methods used to measure pressure, temperature, ion exchange, and voltage, nevertheless converged to license the full ontological status of molecules.

In cognitive psychology, prototype studies in the 1970s revealed just such a robust phenomenon. When asked to rate how representative an object is (*e.g.*, robin, chicken, etc.) of a certain class (*e.g.*, bird), subjects' performance was the same on both ranking and reaction time tasks (Rosch [1973]; Rosch and Mervis, Catlin and Rosch [1976]; Mervis and Rosch [1981]). The convergent results of these diverse test methods are taken by psychologists to indicate that the prototype effects represents a real (non-artifactual) feature of mental organization. Similar convergence of measurements can be found in personality and clinical psychology, where diverse scales are used to identify the same personality traits, and predict the same performance (Norman [1963], pp. 374–91).

Another 'test' routinely deployed in the appraisal of an experimental finding is the extent of integration of that effect into a body of information that is already secure. There seems to be the following methodological principle at work here: Typically, any single effect is more likely to be real if it coheres with, as opposed to contradicts or undermines, well-established findings. Add to this methodological principle a widely held rule of confirmation: A theory is supported to the extent that the evidence favors it over plausible rivals. This rule of confirmation, when applied to explanatory integration, recommends that findings most appropriately united with the most plausible theory are also the most likely to be real or genuinely significant; integrative survival is diagnostic of authenticity.

## 5 SUBMISSION AND EDITORIAL BIASES

Although Faust has charged that psychologists suffer an availability bias, causing them to overlook the frequency of spurious significant findings, it is worth noticing that this is a substantial hypothesis which still awaits testing. It is not entirely clear that Faust, or the multitude of other students of human reasoning, are claiming that the availability bias is a clear example of human irrationality. But they do regard it as a sufficiently serious error to warrant

<sup>11</sup> The classic description of this procedure is Campbell and Fiske [1959]. Wimsatt (1981) is a clear and thorough discussion of the robustness of theoretical objects, events, properties, processes and states, and the emergence of robust objects under different forms of methodological scrutiny and manipulation. That paper also addresses the threat that, despite the experimentalist's honest efforts, tacitly similar presuppositions of diverse methods conspire to produce converging results, a phenomenon called 'bias amplification' or 'pseudo-robustness'.

normative recommendations for improvement. In the case of editorial and submission biases, Faust claims the root problem is availability:

The end result is that individuals apply individual judgement guidelines, few or any of which are likely to exploit maximally the valuable information potentially available through knowledge of base rates, and many of which may be in serious error. . . . They [scientists] could also extend current information about the frequency of unpublished nonsignificant results and its bearing on the likelihood of making false positive errors when interpreting experimental results. One might study scientist's estimates, under commonly encountered conditions, of the base rates for obtaining spurious significant results and determine the extent to which this shapes their conclusions when performing such judgment tasks as evaluating the merits of specific research results ([1984], p. 94).

There are two general problems with this prescription, however, one in estimating the base rates in the first place, and another in knowing what to do with the base rates once you have them. In order to determine the base rate for a certain effect reaching significance, we would first have to estimate the total number of studies, both published and unpublished, that have investigated a particular effect. We could then determine the relative frequency with which the effect reaches significance. But if we are to determine the relative frequency of spurious significant results, obviously we must be able to correctly identify particular significant results as spurious. How might we do this?

In Section 4, I described a widely held assumption of experimental methodology (the principle of robustness) according to which authentically significant effects should be reproducible by similar methods, measurable by diverse procedures and, over time, well integrated into a theory's best models. Guided by this principle, we might be able to identify spurious significant results by observing which effects are robust—those results reaching significance that are reproduced and survive integration—and which do not. Therefore, by searching the literature, we can use this diagnostic test for the presence of real phenomena to determine the relative frequency with which spurious significant results are published.

The need to consider base rates, however, was said to derive from the tendency of professional journals to accept papers with results identified as significant. It is in these journals that, over time, real effects survive and prosper, and spurious results initially reported as significant give way under the stress of fruitless attempts either to replicate and extend the effect, or to use the putative effect as a tool in the testing of other phenomena. Those who criticize the policy of neglecting this base-rate measure, then, are in a peculiar dialectical position. This policy, they claim, is in serious error. Yet, they distinguish spurious from authentic significant findings by deploying standards of reliability employed in the very journals whose neglect of base rates is in dispute. If this is correct, there may be very little to be gained by estimating base rates. Over time, the principle of robustness and the import of integrative survival work against the perennial commission of Type I errors.

The critic of significance testing might object that just because a reported effect disappears or fails integration in future literature does not mean that the reported effect is spurious; these factors are merely diagnostic of actual insignificance, not definitive of it. But, of course, because our science is currently incomplete—we don't yet know how the story turns out—any principles of theory choice will carry with them some epistemic risk, no matter how small. Such principles are ampliative, even if not especially speculative. The critic must therefore tolerate at least some degree of uncertainty in the methods of theory evaluation and choice, or in the standards for distinguishing real from spurious effects, on pain of requiring that researchers be able to correctly predict the future history of science.

## 6 CONCLUSION

If we are to render intelligible the researcher's behavior, we must attribute to her a belief in the robustness of psychological phenomena; otherwise, attempts at replication, extension, triangulation, and a host of other experimental aims, appear either mysterious or unreasonable by her own lights. The psychologists' doxastic commitment to the robustness of psychological phenomena is a special instance of the more general scientific confidence that the robustness of real phenomena will, in the long run and given appropriate background assumptions, guide a reliable (even if imperfect) methodology to the identification of approximately true theories.

Our confidence in the methodology of significance testing and literature review has been vindicated in perceptual and cognitive psychology, disciplines that have enjoyed remarkable progress and consensus in this century, and whose successes are none the less deeply dependent on reigning standards of significance testing. This theoretical success in cognitive and perceptual psychology would be puzzling if significance testing is, as Meehl claims, 'a poor way of doing science'. Criticisms of significance testing are typically framed in purely statistical terms, and levelled without appeal to the role of substantial theory in guiding statistical method. Here as elsewhere in science, whether or not a method works depends as much on the character of the world as it does on the intentions and outlook of the investigator. Although normative refinements can yield welcome improvements in methodology, the world, too, is a sobering and corrective influence on theory construction.<sup>12</sup>

*Loyola University  
of Chicago*

<sup>12</sup> In a work recently completed [Trout, unpublished], I argue that striking and brisk theoretical progress in the behavioral and social sciences is owed to the introduction of quantitative (mainly statistical) methods which rescued those disciplines from the frailties of earlier narrative approaches. According to the 'measured realism' advanced there, the successful application of those quantitative methods requires an explanation that confers on the (observable *and* unobservable) world a corrective influence on rationality.

## REFERENCES

- ARKES, H. and HAMMOND, K. (eds.) [1986]: *Judgment and Decision Making*. Cambridge: Cambridge University Press.
- CAMPBELL, D. T. and FISKE, D. W. [1959]: 'Convergent and Discriminant Validation by the Multitrait–Multimethod Matrix', *Psychological Bulletin*, 56, pp. 81–105.
- CHAPMAN, L. J. and CHAPMAN, J. P. [1969]: 'Illusory Correlation as an Obstacle to the Use of Valid Psychodiagnostic Signs', *Journal of Abnormal Psychology*, 74, pp 271–80.
- CHERNIAK, C. [1986]: *Minimal Rationality*. Cambridge, Mass.: MIT Press/Bradford Books.
- COHEN, L. J. [1986]: *The Dialogue of Reason*. Oxford: Oxford University Press.
- DAVIDSON, D. [1976]: 'Hempel on Explaining Action', *Erkenntnis*, 10, pp 239–53; reprinted in D. Davidson [1980]: *Essays on Actions and Events*, pp. 261–75. Oxford: Oxford University Press.
- FAUST, D. [1984]: *The Limits of Scientific Reasoning*. Minneapolis, Minn.: University of Minnesota Press.
- GIERE, R. [1988]: *Explaining Science*. Chicago: University of Chicago Press.
- GOLDMAN, A. [1986]: *Epistemology and Cognition*. Cambridge, Mass.: Harvard University Press.
- GREEN, B. F. and HALL, J. A. [1984]: 'Quantitative Methods for Literature Reviews', *Annual Review of Psychology*, 35, pp. 37–53.
- HEDGES, L. [1987]: 'How Hard is Hard Science, How Soft is Soft Science?', *American Psychologist*, 42, pp. 443–55.
- HUNTER, J. E. and SCHMIDT, F. L. (1990): *Methods of Meta-Analysis: Correcting Bias in Research Findings*. Newbury Park, CA: Sage Publications.
- KAHNEMAN, D. and TVERSKY, A. [1973]: 'On the Psychology of Prediction', *Psychological Review*, 80, pp. 237–51.
- LUGG, A. [1987]: 'Review of *The Limits of Scientific Reasoning*', *Philosophy of Science*, 54, pp. 137–8.
- LYKKEN, D. T. [1968]: 'Statistical Significance in Psychological Research', *Psychological Bulletin*, 70, pp. 151–9; reprinted in D. E. Morrison and R. E. Henkel (eds.) [1970], pp. 267–79.
- MEEHL, P. E. [1954]: *Clinical versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*. Minneapolis, Minn.: University of Minnesota Press.
- MEEHL, P. E. [1967]: 'Theory Testing in Psychology and Physics: A Methodological Paradox', *Philosophy of Science*, 34, pp. 103–15; reprinted in D. E. Morrison and R. E. Henkel (eds.) [1970].
- MEEHL, P. E. [1978]: 'Theoretical Risks and Tabular Asterisks: Sir Karl, Sir Ronald, and the Slow Progress of Soft Psychology', *Journal of Consulting and Clinical Psychology*, 46, pp. 806–34.
- MEEHL, P. E. [1985]: 'What Social Scientists Don't Understand', in D. W. Fiske and R. A. Shweder (eds.), *Metatheory in Social Science: Pluralisms and Subjectivities*, pp. 315–38. Chicago: University of Chicago Press.
- MEEHL, P. E. [1990a]: 'Appraising and Amending Theories: The Strategy of Lakatosian Defense and Two Principles That Warrant It', *Psychological Inquiry*, 1, pp. 108–41.

- MEEHL, P. E. [1990b]: 'Why Summaries of Research on Psychological Theories are Often Uninterpretable', *Psychological Reports*, 66, pp. 195–244.
- MERVIS, C. B., CATLIN, J. and ROSCH, E. [1976]: 'Relationships Among Goodness-of-Example, Category Norms, and Word Frequency', *Bulletin of the Psychonomic Society*, 7, pp. 283–94.
- MERVIS, C. B. and ROSCH, E. [1981]: 'Categorization of Natural Objects', *Annual Review of Psychology*, 32, pp. 89–115.
- MORRISON, D. E. and HENKEL, R. E. (eds.) [1970]: *The Significance Test Controversy*. Chicago: Aldine.
- NORMAN, W. T. [1963]: 'Personality Measurement, Faking, and Detection: An Assessment Method for Use in Personnel Selection', *Journal of Applied Psychology*, 47, pp. 225–41; reprinted in D. N. Jackson and S. Messick (eds.) [1978], 2nd edn, *Problems in Human Assessment*, pp. 374–91. Krieger Publishing.
- PUTNAM, H. [1974]: 'On the "Corroboration" of Theories'; reprinted in H. Putnam [1975], *Philosophical Papers: Volume One*, pp. 250–69. Cambridge: Cambridge University Press.
- ROSCH, E. [1973]: 'On the Internal Structure of Perceptual and Semantic Categories', in T. Moore (ed.), *Cognitive Development and the Acquisition of Language*, pp. 111–44. New York: Academic.
- ROSCH, E. and MERVIS, C. B. [1975]: 'Family Resemblances: Studies in the Internal Structure of Categories', *Cognitive Psychology*, 7, pp. 573–605.
- ROSENTHAL, R. [1969]: 'Interpersonal Expectation', in R. Rosenthal and R. L. Rosnow (eds.), *Artifact in Behavioral Research*, pp. 181–277. New York: Academic.
- ROSENTHAL, R. [1976]: *Experimenter Effects in Behavioral Research* (enlarged edn). New York: Irvington.
- ROSENTHAL, R. [1979]: 'The "File Drawer Problem" and Tolerance for Null Results', *Psychological Bulletin*, 86, pp. 638–41.
- ROSENTHAL, R. and ROSNOW, R. L. [1984]: *Essentials of Behavioral Research*. New York: McGraw-Hill.
- STICH, S. [1990]: *The Fragmentation of Reason*. Cambridge, Mass.: MIT Press/Bradford Books.
- TROUT, J. D., *Measuring the Intentional World*. Unpublished manuscript.
- WASON, P. and JOHNSON-LAIRD, P. [1972]: *Psychology of Reasoning*. Cambridge, Mass.: Harvard University Press.
- WIMSATT, W. C. [1981]: 'Robustness, Reliability, and Overdetermination', in M. B. Brewer and B. E. Collins (eds.), *Scientific Inquiry and the Social Sciences*, pp. 124–63. San Francisco: Jossey-Bass.