MEASURED REALISM AND STATISTICAL INFERENCE: AN EXPLANATION FOR THE FAST PROGRESS OF "HARD PSYCHOLOGY"

Abstract

The use of null hypothesis significance testing (NHST) in psychology has been under sustained attack, despite its reliable use in the notably successful, so-called "hard" areas of psychology, such as perception and cognition. I argue that, in contrast to merely methodological analyses of hypothesis testing (in terms of "test severity", or other confirmation-theoretic notions), only a patently metaphysical position can adequately capture the uneven but undeniable successes of theories in "hard psychology". I contend that Measured Realism satisfies this description, and characterize the role of NHST in hard psychology.

MEASURED REALISM AND STATISTICAL INFERENCE: AN EXPLANATION FOR THE FAST PROGRESS OF "HARD" PSYCHOLOGY

1. Introduction

Contemporary psychology is actually a complex array of different subject matters, encompassing physiological, perceptual, cognitive, personality, social, counselling, clinical, and educational psychology. Each has its own special methods, but most of those methods are statistical. There is much controversy over whether the so-called "soft" areas of psychology -- social, counselling, clinical, and educational -- have made any progress over the last thirty years or so.[1] By contrast, theories in physiological psychology have advanced briskly, but many would argue that the success is due to the salubrious influence of sophisticated theory in biology. Between these extremes resides perception and cognition, two important areas of psychology that have enjoyed substantial progress over the last three decades. We now have complex and detailed accounts of language, integrating evidence from PET scans, ERP studies, and a variety of behavioral measures. We have an elaborately developed body of theory in vision, detailing both lower- and higher-level processes. In cognition, there are detailed accounts of memory (both implicit and explicit), and problem solving. Why have theories in perception and cognition succeeded when using significance testing, while the "softer" areas of psychology have failed?

Two broad approaches have emerged to explain how Null Hypothesis Significance Testing (NHST), the chief method for establishing causal relations in psychology, might contribute to progress in psychology. One is typified by faith in the power of method alone. After all, when confronted with nature's awesome complexity, Pavlov's methodological dictum seems a mortal's only hope: "With a good method, even a rather untalented person can accomplish much" (quoted in Todes 1997, p.228). In fact, the father of classical conditioning went so far as to claim that "Everything is in the method, in the chances of attaining a steadfast, lasting truth... ." (p.211) This sentiment has dominated methodological reflection in the behavioral and social sciences in the twentieth century. Indeed, it is a resilient legacy of positivism that theories are treated not as

descriptions of causal relations of reality but rather as mere instruments for the testing of hypotheses. This foundational view has been abandoned in every successful disciplinary area of psychology, and its rejection is evident in the structure of the theories themselves. Models of internal processing have been proposed, and they are assessed in terms of their "psychological reality". So it is especially surprising that a Pavlovian, theoretically denuded conception of method operates nowhere more forcefully than in the so-called "significance test controversy". The roster of participants in this controversy reads like a guest list of luminaries and pioneers in 20th century psychology and statistics: Paul Meehl, Robert Abelson, Jacob Cohen, David Lykken, Robert Rosenthal, and John Tukey.[2]

According to this instrumentalist conception of method, the advance of a theory is credited to a merely methodological fact that the hypothesis passed a test that was "severe", or that the inference was “strong”. But there is another view, one which focuses not just on method but on the quality of the theory in light of which method is applied. Reflecting upon scientific progress, the scientific realist's metaphysical assertion seems the only explanation for mortal success. According to this scientific realist view, scientific progress is explained first in terms of the approximate truth of the theories to (or in light of) which an appropriate methodology is applied. In consequence, virtually all progress is theory-dependent.

One thing is clear: Whatever their level of success, perceptual and cognitive psychology are not as successful as contemporary physics, and so do not deserve the success-explanations accorded physical theories. In the case of a mature science such as physics, "the actual theoretical tradition has an epistemically privileged position in the assessment of empirical evidence. Thus, a 'total science' whose theoretical content is significantly in conflict with the received theoretical tradition is, for that reason, subject to 'indirect' but perfectly real prima facie disconfirmation relative to an empirically equivalent total science which reflects the existing tradition." (Boyd 1983, p.209) Whatever else can be said in their favor, perceptual and cognitive psychology cannot be said to disconfirm an affiliated theory simply because it is incompatible with them.

One difficulty in explaining the differential success of the hard and soft areas of psychology

is that the dominant mode of discourse among methodologically reflective psychologists is deeply anti-metaphysical, in the spirit of the empiricist Pavlov. Rarely is there reference to a theory, or any of its parts, "accurately describing the causal structure of the world". Instead, successful theories and their inferences -- ones that are predictively accurate, explanatorily powerful, and theoretically fruitful -- are referred to not as approximately true, but as "strong" (Harlow 1997, p.8). The problem here is an old one. An approximately true theory will also be strong in the psychologist's sense, but it will be much more. The critic of NHST 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.

It is my contention that the failures of NHST are virtually all theory-dependent, and that only a measured scientific realist interpretation of psychological theories can explain the pattern of failures and successes. Let us first turn to the method of NHST.

2. Null Hypothesis Significance Testing

In standard psychological experiments, 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 by chance if two subgroups were drawn again from that population. The standard level of risk or significance is normally set at .05; that is, the null hypothesis can safely be rejected if the risk of duplication due to chance falls below .05. The rejection of the null hypothesis is normally interpreted as confirming the test, "alternative" hypothesis.

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. The errors with which we shall be concerned are Type I. The first complicating factor derives from the simple repetition of experiments on the same dependent variable. An effect which reaches the standard .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 due to random fluctuation in performance, generating results which are spuriously identified as significant, even where there is no real relationship between the independent and dependent variables.

The NHST controversy remains unresolved, but its consequences are enormous. I will argue that this dispute derives from dual demons: the expectation of the triumph of the Pavlovian hope, and the treatment of NHST as a theory-neutral strategy. The result will be that NHST has perfectly acceptable uses, given particular aims. But also it can have a desirable impact in theoretically ambitious domains, when the theories in question are good ones. Indeed, NHST has been incredibly resilient. In the words of one commentator: "Given the many attacks on it, null hypothesis testing should be dead." (Rindskopf 1997, p.319) But some critics of NHST defend it in specific contexts. Meehl contends that significance tests are useful in at least four circumstances:

1. In early stages of research, to help decide whether a trend is worth pursuing with a larger *N*, improved measures, varied contexts, or is more plausibly a "chance" (=random error, not nuisance factors!) difference.

2. In technological contexts; is intervention A stronger than B (even if weakly so)? E.g., efficacy of two antibiotics against a grave infection.

3. In theoretical contexts, when a theory *or its sole competitor* is strong enough to entail *no effect*, so that H_0-refutation can be a strong falsifier. E.g., fifth force, ESP, noncognitive learning theories.

4. In theoretical contexts, when a theory is strong enough to derive a numerical point value *H** or narrow interval (*H** Å _*H**). Here the "null" is not the numerically null H_0, and the strong use can falsify, hence can corroborate, given high power and *failure* to falsify (Karl Pearson's original use of chi-square). (P. Meehl, personal communication, Sept. 12, 1996)

The first two contexts evidently don't depend upon theories that are approximately true, specific, and richly structured. In such atheoretical contexts, NHST can be used as a crude but effective tool when the only question is whether the null is false. This occurs in particular when we are not interested in achieving theoretical understanding, but rather only in deciding what course of action to take. There is, of course, a separate question whether the contexts that might instantiate (1) and (2) are genuinely atheoretical, or whether their theoretical character are concealed, as is so often the case, in familiar instrumentation and background knowledge.

Let us grant, for the sake of argument, that NHST proceeds atheoretically in contexts (1) and (2). It is the success of NHST in contexts (3) and (4) that express the sentiment of this paper. While Meehl prefers to dub such theories "strong", I regard them as "sufficiently approximately true". Once cast in metaphysical terms, the content of (3) and (4) therefore depart from mainstream cautions about the use of NHST. According to this view, if a theory and its auxiliaries are good enough, they can constrain the interpretation of results when either the null is supported or when it is rejected. Distinguishing between the two indicates a test's (low or high) severity. Identifying a test as severe, like saying that it is "strong", is a way of proposing a theory-neutral, merely methodological standard of a good test. But as we shall see, the goodness of a test is a theoretical

property. In order to illustrate the difference between the misleadingly modest description of a testing context and a more accurately depicted realist one, let us consider a case in which a testing context presupposes theoretical constraints -- a discussion of "test severity".

In the philosophy of science literature, there has long been an implicit assumption that some tests of hypotheses are better than others. A test which is insensitive to the target variable will fail to detect an effect due to treatment. Some tests will yield positive results even when the treatment is ineffective. It is this feature that various philosophers have attempted to capture in principles and conditions the provision of good tests of and good evidence for a hypothesis. We can find such accounts of a "severe test" in the work of Popper and Horwich. Deborah Mayo advances a conception of test severity which is characteristic of the antimetaphysical, or merely methodological, inclination of current confirmation theory. Mayo's version draws upon the statistical theory of error:

> (**SC**) There is a very high probability that test *T* would not yield such a passing result, if *h* is false [footnote deleted].
>
> *A test's severity is one minus the probability it yields some such passing result (or other), given the hypothesis passed is false.*
> (Mayo 1991, pp.529-531)

Notice that, although the present account of severity is elucidated in statistical terms, the hypotheses tested needn't themselves be statistical.[3]an isolated hypothesis to the control of an experiment, but only a whole group of hypotheses. When experiment is in disagreement with his predictions, it teaches him that one at least of the hypotheses that constitute this group is wrong and must be modified. But the experiment does not show him the one that must be changed." (Duhem 1954, p.185)

In light of the fact that credit for a good test must be shared with the auxiliary hypotheses, contingent developments in psychology made it possible to assign credit accurately. Many good tests in paleontology (using radiometric dating) depended upon advances in physics. For the first time, there was a secure auxiliary hypothesis. Similarly for the relation between genetics and evolutionary theory, and more pertinent to our current purposes, the relation between signal processing and psycholinguistics. In such cases, the auxiliary theory has been tested independently of its reliable role as an auxiliary.

There is only one plausible explanation for the fast progress of hard psychology: The best current perceptual and cognitive theories possess substantial descriptive components that are approximately accurate characterizations of unobservable phenomena. But this realist explanation for the fast progress of hard psychology must honor certain descriptive facts about the uneven history and contours of perceptual and cognitive psychology.

3. Saying Less than You Could

In the last section, we saw how philosophical elucidations of test conditions masquerade as theoretically neutral, but whose reliability demands a metaphysical interpretation when the scientific details are considered. We will now examine an antimetaphysical strategy, similar to the expressions of method found in the NHST controversy, which attempts to identify a necessary condition of a good test. It is commonly supposed that a good theoretical test of a hypothesis is one which the probability of the hypothesis is low according to any but theory under test. After all, true outcomes based on multiple predictions come cheap (Dawes 1988, Gilovich 1991, and Piatelli-Palmarini 1994). As Giere puts it: "Relative to everything else known at the time (excluding the hypothesis being tested), it must be improbable that the prediction will turn out to be true." (Giere 1984, p.103) We should understand Giere's condition on a good test, outlined above, as a regulative ideal: As much as possible, we should not use the (probabilistic) justification we may already have for the hypothesis in designing a new test of that hypothesis. In order to protect against the trivialization of testing conditions easily produced by the simple derivation of the

hypothesis from background and initial conditions, we need to formulate some appropriate type of restriction barring vacuous confirmation.

But notice that, should the prediction fail to meet test standards due to its over-generality, the present approach presents the hypothesis as infirm for methodological reasons. The idea that our personalities are determined by planetary positions at our birth should be rejected not because it fails a methodological standard, but because it is metaphysically defective. The methodological criticism is certainly accurate, but it is far too weak. To cite another case, it is true that a test of clairvoyance may be suspect because it violates the low probability requirement, but surely it is suspect for substantial, metaphysical reasons also: it postulates the existence of causal mechanisms that are inconsistent with established theoretical evidence. Now, while merely methodological standards of confirmation are pleasingly atheoretical, their ontological squeamishness leaves interpretations of the data far too unconstrained; when test conditions fail, they do so because the theory insufficiently constrains the ranges of values that the variables can take.

Hypothesis testing strategies in any science isolate a hypothesis for examination. But in order for us to determine the genuine truth (or approximate truth) of the test hypothesis, we must be confident that the truth of the hypothesis must be improbable on the (assumed) falsity of the theoretical commitments used in arranging the test of the hypothesis. This assurance is difficult to deliver in every area of psychology. Often the theories supplying the hypotheses to be tested are not very good theories -- whether we call the theories weak or false. In order to reliably infer the alternative upon rejection of the null, it must be the case that the values of the related variables are roughly accurate. The only way to establish those values reliably is to use auxiliary theories that are relatively independent of the hypothesis under test. And although the relevant notion of independence may be elucidated in statistical terms, it cannot be established without metaphysical commitment. Measured Realism is able to capture the evidence of independence, and thus to explain how NHST has helped perceptual and cognitive psychology progress where in fact they have.

4. Measured Realism and NHST

I want to close with some suggestions about how the successful marriage of current perceptual and cognitive theories with NHST can be best explained on the view I have called Measured Realism. If sections 2 and 3 indicate that "merely methodological" accounts of hypothesis-testing cannot adequately explain why NHST works when it does, we can now examine the way in which the Measured Realist alternative might account for uneven success of NHST in the hard areas of perception and cognition, and its failure in soft areas such as personality, counselling, etc. The answer, at this point, has taken shape. Approximately true theories, or the relevant parts of them, make for sensitive tests. Such theories afford clear statements of independence, thus mitigating the chief concern about NHST. By contrast, a poor theory that does not specify the relations among its parts (and thus renders vague predictions) is likely to turn up significant effects, precisely because the test is so weak, so low in severity. Paul Meehl (1997) makes this point by comparing the vague, low-risk hypothesis that it will rain sometime in April, to the specific, risky hypothesis that it will rain 10 of the days in the first half of April. Independence, then, is a source of constraints, making the hypotheses sufficiently specific and risky to be informative.

The contingencies of the history and subject matter of social and behavioral theories have made them too weak and undeveloped to aid reliably in the improvement of affiliated theories; progress has indeed been uneven. Imagine someone—following the realist argument about core areas in physics—trying to defend the claim that only the approximate truth of some sociological account could explain reliable applications of some particular quantitative method. Rather, there are lots of other plausible explanations for this apparent reliability. Maybe the point interval prediction is too weak. Maybe the hypothesis is too vague. Maybe the standard of significance is too lax. Therefore, although there is adequate evidence for a realist interpretation of the social and behavioral sciences, it is a more modest, measured realism than that warranted by the natural and physical sciences. The Measured Realism I defend includes four theses:

1. At least some of the quantitative methods and practices of science, including NHST,

reliably detect and identify some of the central posited entities, processes, states, properties and events in the social and behavioral sciences.

2. Sometimes behavioral and social scientific theories, more often the generalizations that constitute them, are confirmed by the application of accepted methodological standards in the field. This process of confirmation is most evident in such activities as the refinement of measurement procedures and in the successful introduction of controls.

3. In our best psychological and social theories, confirmation relations important to theoretical progress in those sciences are epistemically bite-sized and, accordingly, some parts of the tested theory are typically confirmationally adipose or extraneous. The methodological independence (or "diverse testability") of theoretical objects, processes, events, states and properties, provides evidence that laws about those theoretical kinds are relatively detachable from theories, and so confirmable relatively independent of the particular theories in which those laws were expressed.

4. The reality described by our best psychological and social theories is independent of us *as* observers of, or *as* thinkers about, that reality.

Measured Realism is weaker than the robust realism set out at the beginning of the paper. While robust realism can explain the remarkable success of the mature sciences of, say, physics (on the assumption of its approximate truth), Measured Realism provides the best explanation for the modest theoretical success of the social and behavioral sciences, particularly the success enjoyed upon the introduction of quantitative methods. As an expression of scientific realism, measured realism may appear feeble and hedged. Gone is the familiar commitment to the approximate truth of background theories and the unqualified appeal to the reference of theoretical terms. But in light of the fact that Measured Realism is designed to be appropriate to the social and behavioral sciences,

this modesty is as it should be; the social and behavioral sciences are demonstrably less mature than the physical and biological sciences for which the more daring version of scientific realism was originally and suitably proposed. According to Measured Realism, we have evidence for the approximate truth of at least some laws, whereas austere realism is committed only to the existence of dispositions explaining an observed distribution, and entity realism only to the existence of theoretical entities, but not to the approximate truth of laws and theories.

The most important concept in confirmation theory in determining goodness of test concerns the independence of hypotheses. The relevant sense of independence, however, is more difficult to determine in the operation of immature sciences—such as current cognitive science and late 19th century electrodynamics—than in mature sciences, such as contemporary electrodynamics. Since we now understand the theoretical picture in physics in much greater detail, we know which potential differences it is safe to abstract from, and thus the relevant sense in which one part of the model (such as the unit charge) can be said to be independent of another (such as gravity). In his discussion of Perrin's various independent derivations of Avogadro's number, Peter Kosso says:

> "Perrin measured the same physical quantity in a variety of different ways, thereby invoking a variety of different auxiliary theories. And the reason that Perrin's results are so believable, and that they provide good reason to believe in the actual existence of molecules, is that he used a variety of independent theories and techniques and got them to agree on the answer. The chances of these independent theories all independently manufacturing the same fictitious result is small enough to be rationally discounted." (1989, p.247)

The idea is that the fictitious confluence of results under apparently independent test is rationally negligible, leading to the belief that surprising results confer special confirmation on a theory so long as those results can be explained on the theory in an intellectually satisfying way. Here are a

few examples.

Measured Realism dictates that many of the laws expressing theoretical relations are epistemically bite-sized. The route to an epistemology of such local scale is through the notion of disengaged or adipose theory. For any two measurement procedures applied to the same quantity, there are bits of associated theory in virtue of which the tests or procedures are different. Although any single measurement procedure may engage every piece of associated theory from which it is drawn—and so there is no adipose theory for any single measurement procedure—the fact that two (or three or four) procedures can isolate the same quantity demonstrates that the measurement of that quantity does not depend on any particular theory. If so, then its value can be established independently of any particular theory. It is this independence from a particular theory that supports Measured Realism, for it shows that we can hold measurement outcomes of a selected science to be cumulative without necessarily being approximately true, and it requires that some of the theoretical knowledge acquired concerns bits of theory larger than the theoretical term.

There is much more of this story to tell, and it is told elsewhere (Trout 1998). The independence required to formulate specific hypotheses (both null and alternative) is evident in many features of research in perception and cognition: diverse testing, confirmation of contemporary hypotheses by archaic evidence, the occasional autonomy of experiment over theory, and the successful mercenary reliance of one scientist on an expert in a remote domain. While specific perceptual and cognitive theories are not so secure that theories that conflict with them are thereby disconfirmed, neither are they awash in vagueness and confusion. NHST certainly has its suspect applications, but these frailties are based in theory, not method. Historical contingency distributes its theoretical virtues unevenly, and while a good method helps, no amount of methodological cleverness can compensate for a bad theory.

References

Boyd, R. (1983) "On the Current Status of Scientific Realism", in R. Boyd, P. Gasper, and J.D. Trout (eds.), The Philosophy of Science. Cambridge, MA: MIT Press, pp.195-222.

Dawes, R. (1988), Rational Choice in an Uncertain World. New York: Harcourt, Brace, Jovanovich.

Duhem, P. (1954), The Aim and Structure of Physical Theory. Translated by P. Wiener. New York: Atheneum.

Giere, R. (1984), Understanding Scientific Reasoning (2nd ed.). New York: Holt, Rinehart & Winston.

Gilovich, T. (1991), How We Know What Isn't So. New York: Free Press.

Harlow, L., Mulaik, S., Steiger, J. (eds.), What If There Were No Significance Tests?. Mahwah, NJ: Lawrence Erlbaum Associates.

Kosso, P. (1989), "Science and Objectivity", The Journal of Philosophy, 86, 245-257.

Loftus, G. (1996), "Psychology Will be a Much Better Science When We Change the Way We Analyze Data", Psychological Science, vol. 5(6), 161-171.

Meehl, P. (1997), "The Problem is Epistemology, Not Statistics: Replace Significance Tests by Confidence Intervals and Quantify Accuracy of Risky Numerical Predictions", in L. Harlow, S. Mulaik, J. Steiger, (eds.), What If There Were No Significance Tests?. Mahwah, NJ: Lawrence

Erlbaum Associates, pp.393-425.

Piattelli-Palmarini, P. (1994), Inevitable Illusions. New York: Wiley.

Rindskopf, D. (1997), "Testing 'Small', Not Null, Hypotheses: Classical and Bayesian Approaches", in L. Harlow, S. Mulaik, J. Steiger, eds., What If There Were No Significance Tests? Mahwah, NJ: Lawrence Erlbaum Associates, pp.319-332.

Todes, D. T. (1997), "Pavlov's Physiology Factory", Isis, 88, 205-246.

Trout, J.D. (1998). Measuring the Intentional World: Realism, Naturalism, and Quantitative Methods in the Behavioral Sciences. New York: Oxford University Press.

Footnotes

[1] It is worth mentioning that most psychologists believe that progress is possible, and sometimes actual. The jaded critic, one who casts around for dramatic claims of "crisis" in psychology, is in the minority.

[2] The practices of NHST are so inveterate a part of psychology that, when called into question (Harlow 1997), the issue of its integrity demanded systematic examination. Opinions about NHST are so divided that there is now an ad hoc committee assembled to address the question of whether American Psychological Association journal editors should forbid authors to report the results of significance tests, with distinguished outside expert consultants such as Lee Cronbach, Paul Meehl, Frederick Mosteller, and John Tukey.

[3] This fact is quite general. Statistical measures are used to elucidate the notion of independence in confirmation-theoretic approaches, such as high probability or surprise accounts, but these are best understood as expressions of metaphysical independence of entities and laws.