ELSEVIER

Acta Psychologica 108 (2001) 303–307

acta psychologica

www.elsevier.com/locate/actpsy

Book review

Hypothesis Testing Behaviour. Fenna H. Poletiek (2001). Hove, East Sussex, UK: Psychology Press

Hypothesis testing is a controversial theme in several areas of inquiry. In the philosophy of science, the process of conceptualizing hypothesis testing and its role in science created one of the major philosophical discussions of the 20th century, namely that between the verificationism of the logical positivists and Popper's falsificationism. In statistics, ever since Fisher proposed the null-hypothesis testing procedure, there has been an ongoing debate on the issue of hypothesis testing between proponents of different schools, advocating perspectives based on decision theory (i.e., the Neyman–Pearson procedure), Bayesian theories of belief revision, and information theory. Finally, in the behavioural sciences, the experiments carried out by Peter Wason in the 1960s generated 40 years of research into the psychological processes involved in testing behaviour, centered around the question whether people actually possess a tendency to confirm, rather than disconfirm, their theories – the so-called confirmation bias. In her book, Hypothesis Testing Behaviour, Fenna Poletiek takes on the ambitious task of connecting the issues raised in these fields. She does this by interpreting the various stances that have been taken with respect to hypothesis testing, and by reviewing the available evidence. In addition, the monograph contains some interesting new perspectives on old disputes.

The book is composed of four parts. Chapter 1 covers issues on hypothesis testing in the philosophy of science. Chapter 2 contains a selected overview of the major schools of thought in statistical inference. Chapters 3, 4, and 5 are concerned with the empirical study of human behaviour in various testing situations. Finally, in Chapter 6 Poletiek organizes the material of the previous chapters into her own perspective, the probability value model.

Chapter 1 deals with issues in the philosophy of science. Poletiek contrasts verificationist and falsificationist standards of testing to reach a surprising result: confirmatory behaviour is falsifying behaviour. Her argument is based on showing that the verificationist concept of the relevance ratio [1] is equivalent to the falsificationist concept of test severity. [2] The method of proof is based on a Bayesian interpretation

[1] The relevance ratio is defined as $P(H|x,b)/P(H|b)$, where H is the hypothesis of interest, x represents the empirical data, and b is the available background knowledge. Thus, relevance indicates the degree to which the data increase the probability of the hypothesis.

[2] Test severity is defined as $P(x|H,b)/P(x|b)$. Thus, test severity indicates the degree to which the hypothesis increases the probability of the data, or, alternatively, the degree to which the hypothesis is necessary for predicting the data.

0001-6918/01/$ - see front matter
PII: S0001-6918(01)00044-0

of these concepts. Roughly, Poletiek shows that, upon entering the probability of a hypothesis, given the evidence, into Bayes' theorem, algebraic manipulation reveals that the relevance ratio and the severity of a test are necessarily equal. Thus, maximizing relevance and maximizing severity prescribe the same testing behaviour. Although several reservations may be put forward with respect to the method of proof and the interpretation of its result, the discussion shows clearly that verificationism and falsificationism are not necessarily opposed in terms of the behavioural standard they advocate. This is an interesting finding, because it automatically renders the term 'confirmation bias' open to criticism: if verificationism and falsificationism, at least upon one interpretation, result in the same behavioural standard, where does the confirmation bias come from? The second issue on which Poletiek focuses is the asymmetry between test severity and the value of falsifications. She aims to show that, as a test becomes more severe, refuting evidence becomes weaker and confirmatory evidence stronger. Conversely, if a test is less severe, confirmatory evidence is weaker and refuting evidence stronger. These two conclusions, i.e., the equivalence of verificationism and falsificationism in terms of behavioural standards, and the asymmetry of tests, provide the two general themes that resonate throughout the book.

The theme of asymmetrical tests is taken up in Chapter 2, which deals with statistical theories of hypothesis testing. Poletiek describes several statistical approaches to the testing problem: Fisher's null-hypothesis testing procedure, the decision theoretical framework proposed by Neyman and Pearson, the Bayesian approach, and procedures based on information theory. Poletiek shows how the asymmetry between test severity and the value of refutations translates into statistical properties of tests. The asymmetry boils down to the conjecture that, if a hypothesis prescribes that the probability of a specific event is very large, say 0.99, then actually observing the event provides less information than not observing the event: the test is a better falsifier than it is a confirmer. This is nicely illustrated using concepts from information theory: The information provided by a confirmatory or refuting observation is not independent of the respective probabilities of the observation, conditional on the hypothesis, and conditional on relevant alternative hypotheses. This, of course, posits a further problem for the consistent definition of confirmation bias, because it shows that, in a statistical framework, falsifications do not necessarily provide more information than confirmations. The issue of whether confirmation bias can be consistently defined, and, if so, whether people actually display it, is taken up in the next three chapters.

Chapter 3 deals with Wason's famous rule discovery task. The chapter provides an extensive review of the empirical work that has been carried out and the numerous modifications of the test that have been proposed. And, again, the delicate problem of defining the concept of confirmation bias plays an important role. In a thorough discussion of reformulations of the 'tendency to confirm' in terms of positive and negative testing, and necessity versus sufficiency tests, Wason's original claims are stripped down to a point where the conclusion that people actually display a confirmation bias must be regarded as too strong. Rather, the available evidence suggests that subjects tend to use several different strategies in different stages of

inquiry. For example, at the point where the hypothesis is regarded as very uncertain, positive testing (i.e., testing by means of proposing examples that are consistent with the hypothesis) is important, whereas negative testing (i.e., testing by means of proposing examples that are inconsistent with the hypothesis) becomes important once the subject has achieved a reasonable amount of confirmation for his or her hypothesis. Also, Poletiek reviews evidence suggesting that consistently pursuing refutations by negative testing does not help in finding the true rule, which casts further doubt on the adequacy of falsificationism as a behavioural standard.

In Chapter 4, similar conclusions are reached for Wason's selection task. This task generated an even larger body of research, and in addition led to the proposal of several statistical models for participant behaviour. Poletiek discusses the empirical evidence as well as the plausibility of the explanations proposed for it. Included in the discussion are explanations based on defective truth tables, deontic reasoning, mental models, cognitive availability, relevance, and relative set sizes. Also, several statistical inference approaches are discussed and evaluated with respect to their psychological plausibility. At the end of the chapter, Poletiek suggests that many of the proposed explanations can be subsumed under the general concept of relevance. That is, the participant chooses the option that he or she regards as most relevant for the hypothesis. The major benefit of the research based on the selection task, according to Poletiek, is that it has produced insight into which conditions actually render evidence subjectively relevant. Finally, Poletiek illustrates how explanations based on, for example, matching and deontic reasoning may be translated into an explanation in terms of relevance.

At this point, the reader has been informed that falsifying and confirmatory behaviour amount to the same thing, that falsifications are not necessarily the most informative pieces of evidence, and that the actual behaviour of participants is mediated by a myriad of minor variations in experimental procedures. So, the rational arguments for falsificationism as a normative standard, and the interpretation of participant behaviour as violating this standard are exposed to serious criticism. It is therefore surprising to find a consistent definition of confirmation bias in Chapter 5, which deals with decision making under uncertainty. In experiments investigating this kind of behaviour, participants have to choose among a number of tests for a given hypothesis. Both the probability of finding a specific observation if the hypothesis is false, and the probability of finding it if the hypothesis is true, may be available. In this case, confirmation bias can be said to occur if a participant chooses a test that has a higher probability of confirmation than the prior probability of the hypothesis in question. That is, confirmation bias occurs if the participant selects a test that maximizes the probability of confirmation, regardless of whether the hypothesis is true or false. Poletiek reviews some evidence suggesting that this 'real' confirmation bias does indeed occur under specific conditions. However, if the likelihoods of results under the focal hypothesis and some alternative are both given, people are capable of choosing the test that best discriminates between the hypotheses. And, as was the case in the rule discovery and selection tasks, the results obtained vary with minor task variations and can be explained in many ways besides the tendency to protect the favourite hypothesis from falsification.

One such explanation is proposed by Poletiek in Chapter 6, where she explains her own theory of testing behaviour. The probability value model, as it is called, is based on the asymmetrical nature of tests as discussed throughout the book. The model presumes that test choosers base their choice on two characteristics of tests: the probability of obtaining a specific piece of evidence and the value of that piece of evidence. What a particular person maximizes is allowed to vary with person characteristics, such as the extent to which a person believes the theory to be true. For example, a person strongly believing in a theory may opt for a very severe test (minimizing the unconditional probability of confirmatory evidence) which will give a strong confirmation if it yields a result in line with the hypothesis (thus, the value of the confirmatory result is maximized). The strength of the model lies in the incorporation of individual differences into the model. However, as the model is still in a rather sketchy phase, whether it will be a successful approach to human behaviour in testing situations will depend on the outcome of future research.

The general strengths and weaknesses of Poletiek's monograph may be recognized from the above summary. The book has the benefit of being the product of a single author: General themes are consistently maintained throughout the book, and the chapters do not vary widely in style, depth, or technical difficulty. In addition, the scope of the book is impressive: combining philosophy of science, statistical inference, and experimental research into human behaviour is no small task, and Poletiek has succeeded in bringing these viewpoints together. The natural consequence of maintaining the same concepts and distinctions throughout the book, however, is that some expositions are clearly shaped by the view of the author. For example, the alleged equivalence of falsificationism and verificationism is open to criticism, because it depends on a Bayesian interpretation of these philosophies. Certainly, Popper would not have agreed with it for the simple reason that he did not believe one could meaningfully speak of the probability of a hypothesis. A second drawback of the impressive scope of the book is that the author is not equally well informed in each area of inquiry. For example, in the statistical part, the emphasis is on the role of the likelihood ratio in various statistical models. However, Poletiek does not discuss relevant approaches that have explicitly treated the likelihood ratio as the central concept in statistical inference, such as the work of Hacking (1965) on the likelihood principle, and Royall's (1997) monograph on statistical inference in the likelihood paradigm. A final criticism concerns the probability value model. The treatment of the model is rather informal. However, the model could be formalized without substantial problems, for example by introducing person parameters into a decision model, which raises the question why this has not been done. As it stands, the model is too vague to be put to serious (read: severe) tests.

Still, whether or not the reader agrees with Poletiek's analysis of testing behaviour, it is definitely thought provoking. What the book makes especially clear is that the connection between philosophy of science, statistical inference, and psychology, is not straightforward by any means. The problems raised in the process of connecting these disciplines are important and deserve attention in the literature.

References

Hacking, I. (1965). *Logic of statistical inference*. London: Cambridge University Press.
Royall, R. (1997). *Statistical evidence: a likelihood paradigm*. London: Bowman & Hall.

Denny Borsboom
Department of Psychology
University of Amsterdam
Roetersstraat 15, 1018 WB
Amsterdam, Netherlands