UC Merced
Proceedings of the Annual Meeting of the Cognitive Science Society

Title
A Rarity Heuristic for Hypothesis Testing

Permalink
https://escholarship.org/uc/item/8kd818mj

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 22(22)

Authors
Feeney, Aidan
Evans, Jonathan St.B.T.
Venn, Simon

Publication Date
2000

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

A Rarity Heuristic for Hypothesis Testing

Aidan Feeney
Department of Psychology
University of Durham
Science Laboratories
South Road
Durham DH1 3LE
United Kingdom
aidan.feeney@durham.ac.uk

Jonathan St.B.T. Evans & Simon Venn
Centre for Thinking and Language
Department of Psychology
University of Plymouth
Drake's Circus
Plymouth PL4 8AA
United Kingdom
{jevans, svenn}@plymouth.ac.uk

Abstract

This paper presents the results of two experiments designed to investigate the processes underlying the effects of beliefs about probabilities on an hypothesis testing task. Both experiments demonstrate that although such effects exist, they are inflexible in the face of explicit statistical and implicit contextual manipulations of the likely information to be gained from selecting evidence concerning rare features. It is argued that these results suggest the operation of a rarity heuristic in hypothesis testing whilst possible adaptive functions for such a heuristic are discussed.

Probabilities and Hypothesis Testing

Over the last ten or more years, human hypothesis testing, which previously had been viewed as being prone to bias (Wason, 1960; Doherty et al, 1979), has been rehabilitated. One of the central claims to be made during this process of rehabilitation is that human hypothesis testing is in some way adapted to the probabilistic structure of our environment. For example, Klayman and Ha (1987) have argued that confirmation, verification and matching biases, amongst others, may be viewed as the result of a generalised positive test strategy in hypothesis testing. Klayman and Ha demonstrated how such a strategy could be a good heuristic in environments with a realistic probabilistic structure. Their claim is that hypothesis testing tasks such as Wason's 2 4 6 task lead to non-normative behaviour because they encourage participants to adopt a generally sensible strategy in an experimental situation whose probabilistic structure does not match the strategy.

More recently, Oaksford and Chater (1994), in the spirit of Anderson's (1990) more general 'rational analysis' of cognition, have proposed a decision-theoretic account of Wason's selection task (Wason, 1966). This account is based on both the probabilistic structure of the task itself and assumptions about people's understanding of abstract conditional hypotheses. Our aim in this paper is to extend the study of probabilistic effects to another hypothesis testing task and to gain some insight into the mechanisms underlying such effects.

Probabilities and Pseudodiagnosticity

Feeney, Evans and Clibbens (1997) have considered the role of background probabilities in determining performance on the pseudodiagnosticity (PD) task. Pseudodiagnosticity (Doherty et al, 1979) is the tendency to select information relevant to just one of a pair of hypotheses when trying to decide between them. An example of the standard paradigm used to investigate pseudodiagnosticity is taken from Mynatt, Doherty and Dragan (1995):

> Your sister has a car she bought a couple of years ago. It's either a car X or a car Y but you can't remember which. You do remember that her car does over 25 miles per gallon and has not had any major mechanical problems in the two years she's owned it.
>
> You have the following information:
>
> A. 65% of car X's do over 25 miles per gallon.
>
> Three additional pieces of information are also available:
>
> B. The percentage of car Y's that do over 25 miles per gallon.
>
> C. The percentage of car X's that have had no major mechanical problems for the first two years of ownership.
>
> D. The percentage of car Y's that have had no major mechanical problems for the first two years of ownership.
>
> Assuming you could find out only one of these three pieces of information (B, C or D) which would you want in order to help you to decide which car your sister owns? Please circle your answer.

In the standard PD task, as above, an anchor is provided (item A) which provides some potentially supportive evidence for one of the two hypotheses presented in the scenario. This we term the *focal* hypothesis. According to Doherty and his colleagues, the normatively correct answer to this problem is to choose item B which provides - in Bayesian terms - a completed likelihood ratio and allows the diagnosticity of the evidence to be assessed. For example,

we might discover that only 25% of Y's do over 25 mpg, favouring X or that 90% of Y's do over 25 mpg, favouring the Y hypothesis. However, the more common response is for people to choose item C, thus learning more about X. In the study quoted, 28% of participants chose B (deemed correct), 59% chose C and 13% chose D. In the absence of information about Y, however, items A and C provide only pseudodiagnostic evidence for X.

The original interpretation of this apparent error by Doherty et al (1979) - with general support in the later literature - was that it constituted a form of confirmation bias similar to that observed on other tasks such as the Wason 2 4 6 problem (see Evans, 1989, Klayman, 1995 for extended discussion of confirmation bias effects). It is supposed that people think only about the focal hypothesis, fail to consider alternatives and try to find evidence to confirm their favored hypothesis.

However, the analysis of the task becomes more complex if one takes into account background beliefs that the participant may bring to the experiment. Suppose, for example, that you were told that your sister's car had a radio and a top speed of over 165 miles per hour. If the information provided was then that most X's have a radio, according to the standard normative analysis people ought to choose to discover whether most Y's also have a radio. However, since they know a priori that most cars have radios, the participants could reason that this will be true of most Y's as well and that nothing will be learned by choosing this option. On the other hand discovering whether X does over 165 miles an hour (a rare feature among cars) would provide good evidence relative to background beliefs about the likelihood of this feature. Given these beliefs, such evidence could be regarded as being implicitly diagnostic rather than as being pseudodiagnostic. In this case, one can actually argue that the PD choice is correct, because its expected information gain (Oaksford, Chater & Larkin, 1999) or epistemic utility (Evans & Over, 1996) is higher, relative to background beliefs.

Feeney, Evans and Clibbens (1997) have shown that when the initial piece of evidence concerns a rare feature and the second piece concerns a common feature, then people will seek to discover a second piece of evidence about the rare feature, leading to a large drop in the usual PD choice rate. This tendency, to make diagnostic selections when evidence concerning a rare feature is available has been replicated on three different variants of the task (Feeney, Evans and Venn, 2000). In a separate version of the paradigm in which participants rate their degree of belief in the focal hypothesis after one or two pieces of 'pseudodiagnostic' information, we also found (Feeney, Evans and Clibbens, in press) that people are significantly more confident in a hypothesis supported by rare rather than common evidence. These findings support the view that rare information is taken to be implicitly diagnostic.

Whilst the experiments described above have established a robust influence of feature Rarity, it is not clear whether this is due to tacit influence of background beliefs or whether people are consciously reasoning about the expected epistemic utility of the evidence. This ambiguity is indicative of a more general confusion (Oaksford Chater & Larkin, 1999; Klayman and Ha, 1987) in the literature on hypothesis testing where it is unknown whether people's apparent sensitivity to the probabilistic structure of their environment is the result of hard-wired heuristics, or is due to extensive on-line processing of environmental probabilities. We will now describe two experiments designed to resolve this ambiguity.

Experiment 1

In this experiment, we used problems which were structurally identical to those used by Mynatt et al (see above). In Mynatt et al's experiment participants received a scenario containing a target object, said to possess two features, and two hypothesised categories. Next participants received a piece of evidence concerning the rate at which one of the hypothesised categories possessed one of the features. Finally, participants were asked to select one of the remaining three pieces of information to help them make a judgement about category membership.

In Experiment 1 we manipulated the relationship between the rarity of the evidential features and the explicit information presented about the rate at which features were present under either hypothesis. In the belief-compatible conditions, participants were told that a rare feature was present in only 10% of cases for the initial hypothesis (e.g. 10% of car X's do over 165 mph) or that a common feature was present in 80% of cases (e.g. 80% of car X's have radios). In the belief-incompatible conditions, participants were told that the rare feature was present in 80% of X's or that the common feature was present in 10% of X's. If a simple rarity heuristic is operating then we would expect participants still to favour rare features regardless of the explicit information given. However, if they are reasoning on-line about the probability of the evidential features then the percentage data should interact with the Rarity manipulation. Specifically, when told that the common feature is present in only 10% of X's (common feature, belief-incompatible), we might now expect diagnostic choices to go up (and focal choices to be suppressed) even though these involve the common feature. This is because people could reason that most Y's will probably have the feature and hence the choice will be diagnostic. When told that 10% of X's contain the rare feature (belief-compatible) we might also expect a drop in the usual diagnostic choice rates for rare evidence since they will expect Y to have a similar rate. Hence, the on-line processing hypothesis predicts a cross-over interaction between the two variables.

Method

One hundred and eighty seven students from the University of Plymouth took part in this experiment which had a 2x2 between participants design. Each participant received a booklet comprising of an instructions page and four

problems. The basic structure of the problems used was identical to that used by Mynatt et al. The factors manipulated were Rarity and the strength of the initial statistic presented (we will refer to this variable as the Percentage variable). The Rarity manipulation in this experiment was between participants and was achieved by manipulating the first feature about which participants were given some evidence. These features were chosen on the basis of a pre-test and are shown in Table 1.

Table 1: Results of pre-test on materials used in Experiments 1 and 2.

Content	2nd Feature	1st Feature - Common	1st Feature - Rare
House	Garage	Garden	Swimming pool
Engineer	Company car	Earns £14,000 pa	Earns £60,000 pa
Car	Has a radio	Top speed 90 mph +	Top speed 165 mph+
Holiday Villa	Built last 20 years	£150 per week	£1000 per week

The Percentage manipulation was achieved by manipulating, between participant, the strength of the initial piece of evidence. Half of the participants were told that 10% of instances of the focal category shared a feature with the target whilst the other half were told that this figure was 80%. The problem contents employed in this experiment concerned a house, an engineer, a car and a holiday villa. The order of the evidential options was counterbalanced whilst the order of the problems was randomised.

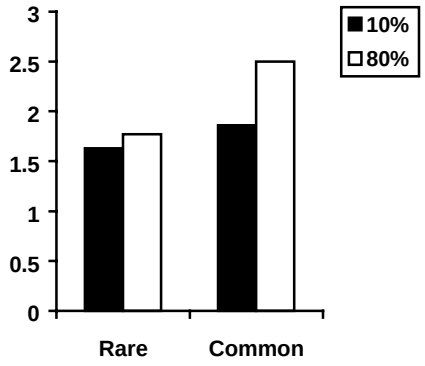
Results and Discussion

Evidence selection patterns, when collapsed across experimental condition, are very similar for all problem contents. On the Engineer problem 41% of selections were of B (diagnostic selections), 50% were of C (further information about the focal hypothesis) and 9% were of D (information for the non-focal hypothesis concerning the second feature). The equivalent statistics for the Spanish Villa problem are 37%, 50% and 13%, 42%, 44% and 14% for the Car problem and 41%, 50% and 9% for the House problem. Selection frequencies for the entire experiment broken down by experimental condition are presented in Table 2. The mean number of pseudodiagnostic, or item C, choices was calculated for each participant across the four problem contents. The mean number of item C choices, broken down by Rarity and Percentage Type, is presented in Figure 1. A 2x2 between-participants Anova was carried out on the mean number of item C choices. A significant main effect was found for Rarity ($F(1, 183) = 4.53$, $MSE = 2.40$, $p < .04$). The mean number of item C choices made by participants in the Rare and Common conditions was 1.70 (S.D. = 1.57) and 2.18 (S.D. = 1.54) respectively. Neither the main effect of Percentage ($F(1, 183) = 2.98$, $MSE = 2.40$, $p > .08$) nor the interaction between Rarity and Percentage ($F(1,183) = 1.19$, $MSE = 2.40$, $p > .25$) were found to be statistically significant.

Table 2: Item choices as a percentage of total choices in each condition for Experiment 1.

	Rare		Common	
	10	80	10	80
Item B	47%	43%	41%	28%
Item C	42%	44%	46%	63%
Item D	11%	13%	13%	9%

Figure 1: Mean Number of Focal Selections by Condition in Experiment 1.



These results suggest the operation of a rarity heuristic in hypothesis testing. Our results contained a significant main effect of Rarity, although the apparent interaction fell short of significance. Whilst it is clear from Table 2 and Figure 1 that the percentage information has no effect on choice rates for rare information, Figure 1 does reveal a marginally significant trend ($p < .06$) for common choices to be debiased in the belief-incompatible condition.

The trend for focal selections to increase when the initial evidence is that less than 50% of focal instances possess a common feature was found previously by Mynatt et al (1993) and interpreted by them as due to the initial evidence disconfirming hypothesis X and switching focus to Y. Although this trend is also consistent with an on-line processing hypothesis, that hypothesis also predicts a corresponding increase in focal choices when rare information was present at 80%. The latter trend was clearly absent. However, the trend which is to be seen in our data is consistent with the claim, made by Mynatt et al, that people select evidence relevant to the hypothesis they believe to be true.

Experiment 2

In our previous experiment we demonstrated that people are relatively insensitive to explicit changes in the initial piece of information that they receive and that the rarity effect is

robust in the presence of such changes. In Experiment 2 we aimed to test whether people are sensitive to contextual changes which should also affect the epistemic utility of their choices.

In this experiment we attempted to reduce the implicit diagnosticity of the rare features by presenting them in a context where they would not be rare. For example, in the Car scenario both types of car were said to be sports cars and hence much more likely to have a high top speed. As all participants in the experiment were told that 80% of Xs possessed either the common or rare features used in previous experiments, the implicit change to the scenarios was the only difference between the two conditions run in this experiment and the 80% conditions of the previous experiment. Combining both sets of conditions gives us a 2x2 design with rarity of initial feature and relationship between alternatives the between subject variables.

Method

One hundred and four new participants from the University of Plymouth were recruited for this experiment each of whom received a booklet comprising of a set of instructions and four problems. The instructions for this experiment were identical to those used in Experiment 1 whilst the problem contents used were almost identical to those used in the previous experiments. The rarity manipulation in this experiment was achieved with the same features as used in previous experiments and all participants were told that 80% of instances of the focal category shared a feature with the target object.

The difference between the new conditions in this experiment and the 80% conditions of the previous experiment is the implicit diagnosticity of the rare feature. Thus, we will refer to the second between participants factor in this experiment as Implicit Diagnosticity.

Results and Discussion

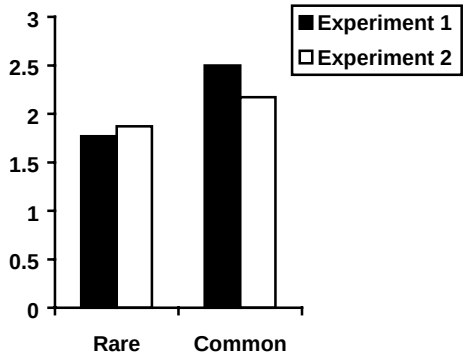
Once again, selection patterns, when collapsed across our new experimental conditions, are very similar for all four problem contents. On the Engineer problem 37% of selections were of B, 50% were of C and 13%% were of D. The equivalent statistics for the Villa problem are 35.5%, 52% and 12.5%, 38%, 51% and 11% for the Car problem and 40%, 49% and 11% for the House problem. Selection frequencies for the entire experiment broken down by experimental condition are presented in Table 3.

Table 3: Item choices as a percentage of total choices in each condition for Experiment 2.

	Low Implicit Diagnosticity (Exp. 2)		High Implicit Diagnosticity (Exp. 1)	
	Rare	Com	Rare	Com
Item B	42%	34%	43%	28%
Item C	47%	54%	44%	63%
Item D	11%	12%	13%	9%

The mean number of item C choices was calculated for each participant across the four problem contents. The mean number of C choices, broken down by Rarity and Implicit Diagnosticity, are presented in Figure 2. In order to examine the effect of our Implicit Diagnosticity manipulation a 2x2 between participants Anova (with Rarity as the second factor) was carried out on the mean number of item C choices made by participants in this experiment and participants in the 80% conditions of Experiment 1. Once again, a significant main effect was found for Rarity ($F(1, 199) = 5.745$, $MSE = 2.365$, $p < .02$). The mean number of item C choices made by participants in the Rare and Common conditions was 1.81 (S.D. = 1.57 and 2.33 (S.D. = 1.55) respectively. Neither the main effect of Implicit Diagnosticity ($F(1, 199) = 0.235$, $MSE = 2.365$, $p > .6$) nor the interaction between Rarity and Implicit Diagnosticity ($F(1,199) = 0.95$, $MSE = 2.248$, $p > .33$) were found to be statistically significant.

Figure 2. Mean number of focal selections in Experiment 2 as a function of condition.



These results demonstrate the persistence of people's tendency to make fewer focal choices when the initial piece of information concerns a rare feature, even when the implicit diagnosticity of that rare feature has been contextually reduced. This provides further evidence of a robust rarity heuristic that is relatively insensitive to contextual variations.

General Discussion

The first conclusion to be drawn from the results of the experiments described in this paper is that they support the findings of Feeney, Evans and Clibbens (1997; in press). People are sensitive to the probabilities of the evidential items about which they reason on the PD task. More important is our failure to moderate the effects of feature rarity using either an explicit statistical manipulation or an

implicit contextual manipulation. The failures of these manipulations suggest that the effect of feature rarity is mediated via a hard-wired heuristic rather than any sophisticated on-line processing of probabilities. Whilst this heuristic is sensitive to rare features of objects, it is insensitive to changes in explicit statistical and implicit contextual information which affect the diagnosticity of those features. Accordingly, although we agree with Oaksford, Chater, and Larkin (1999) who argue that the very existence of probabilistic effects in hypothesis testing tasks indicates that people perform some on-line processing of probabilities, we feel that our results strongly suggest that the extent of such processing is severely limited. The effects of feature rarity on the PD task seem instead to be due to the operation of a relatively inflexible rarity heuristic.

Functional and Dysfunctional Aspects of a Rarity Heuristic

As with any heuristic in judgement or hypothesis testing, a rarity heuristic conveys both advantages and disadvantages. Most obviously, given the results of our experiments, an inflexible rarity heuristic renders the information processor insufficiently sensitive to changes in the diagnosticity of rare experimental features. However, as we have claimed elsewhere (Feeney et al, in press), sensitivity to feature rarity allows us to use our background beliefs about the probability of the evidence to evaluate hypotheses even in the light of normatively incomplete evidence. For example, imagine you have been asked to decide whether your sister's car, which possessed a top speed of over 165 mph and a radio, is a model X or a model Y. Given your background knowledge about the features, you can be more confident that the car is an X when told that 95% of Xs have a top speed of over 165 mph than when told that 95% of Xs have a radio. Thus information about feature rarity may be used to make a decision even when normatively complete evidence is missing.

As well as supporting inference with incomplete information, we believe that another candidate function for a rarity heuristic in hypothesis testing might be checking the limitations of hypotheses. Defining the scope of hypotheses in this way has recently become a topic of interest for cognitive psychologists. For example, Lopez (1995) has found that the majority of participants presented with a premise such as:

Dogs have a merocrine gland 1

and asked if they would prefer to find out whether wolves or bulls had a merocrine gland in order to check the more general premise that

All mammals have a merocrine gland 2

preferred to check bulls rather than wolves. This preference is viewed as being normatively correct as it obeys the notion that the more diverse is the evidence in favour of a hypothesis the stronger the support for that hypothesis is (see Carnap, 1951; Popper, 1959). Osherson, Smith, Wilkie, Lopez and Shafir (1990) have proposed a model of category based induction which captures the diversity principle. This model accounts for people's preference for diverse premises by supposing that the strength of a categorical argument depends on the degree to which the premise categories are similar to both the conclusion category and instances of the lowest level category which includes both premise and conclusion categories.

There are situations in which it may be impossible to make the similarity calculations upon which Osherson et al's model relies. For example, it is common in the literature on category-based induction to use premises with blank predicates i.e. premises about which the participant is unlikely to have any a priori beliefs. This is done to minimize the effects of the predicate on participants' judgements. It is also possible to use blank premise categories in these experiments where the participant had no knowledge about the premise and conclusion categories except their size. In this case although the information required for a similarity calculation is unavailable one can check whether a general hypothesis also applies to a rare or unusual event. Thus we can greatly increase our confidence in the hypothesis (when it can account for the rare event) or limit the hypothesis (when it cannot).

The importance of such a limiting function may be seen when one considers that several lines of theoretical and experimental work suggest that it is the interaction between a heuristic or strategy and the environment in which it is used which determines the success or failure of that heuristic (e.g. Evans, Handley, Harper and Johnson-Laird, 1999; Gigerenzer et al, 1999). This argument was most explicitly made by Klayman and Ha (1987) who defined the probabilistic structure of environments where their positive test strategy would not be successful. The consequences of a mismatch between the environment and the positive test strategy is most dramatically illustrated by the failure of participants on Wason's 2 4 6 task to limit their initial rule thereby leading to a failure to discover the experimenter's more general rule.

The experiments described in this paper demonstrate the use of a heuristic which counteracts the effects of a positive test strategy. The standardly obtained finding on the pseudodiagnosticity task is that subjects tend to search for more information about the hypothesis supported by the existing evidence. In most cases this leads to pseudodiagnostic responding. In our experiments we have demonstrated that the tendency to select information about rare features produces diagnostic responding. In a similar fashion, one can imagine a scientist who is committed to a hypothesis that is too narrow but, because of the use of a positive test heuristic, is unable to find disconfirmation. The attempt to apply this hypothesis to a rare event or phenomena may provide the evidence required for the scientist to broaden the hypothesis.

Whilst we see the diversity principle (Osherson et al, 1990; Lopez, 1995; Spellman, Lopez and Smith, 1999) and a rarity heuristic as being complementary, there is an

important distinction to be drawn between them. Lopez (1995) has argued that the diversity principle does not breach the positive test heuristic. One of the adaptive functions of the rarity heuristic, on the other hand, is that - as in the experiments reported in this paper - it does violate a strategy based on positive testing. Its existence and use in everyday hypothesis testing is likely to be one reason why we are not surrounded by the calamitous results of a reliance on positive testing.

Conclusion

In conclusion, we have demonstrated that the effects of rarity on hypothesis testing, although present, are insensitive to statistical and contextual manipulation. We have argued that these results support the existence of a Rarity heuristic in hypothesis testing. Finally, we claim that if such a heuristic does exist, it is likely, in many cases, to complement the operation of other heuristics known to operate when people select information to help them find out about the world.

References

Anderson, J.R. (1990). The adaptive character of thought. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Carnap, R. (1951). Logical foundations of probability.

Doherty, M.E., Mynatt, C.R., Tweney, R.D., & Schiavo, M.D. (1979). Pseudodiagnosticity. Acta Psychologica, 49, 111-121.

Evans, J.St.B.T. (1989). Bias in human reasoning: Causes and consequences. Hove, UK: Lawrence Erlbaum Associates Ltd.

Evans, J.St.B.T. & Over, D.E. (1996). Rationality and reasoning. Hove, UK: Psychology Press.

Evans, J.St.B.T., Handley, S.J., Harper, C.N.J. & Johnson-Laird, P.N. (1999). Reasoning about necessity and possibility: A test of the mental model theory of deduction. Journal of Experimental Psychology: Learning, Memory and Cognition, 25, 1495-1513.

Feeney, A., Evans, J.St. B.T. & Clibbens, J. (1997). Probabilities, utilities and hypothesis testing. Proceedings of the 19th Annual Conference of the Cognitive Science Society. Mahwah, NJ: Erlbaum.

Feeney, A., Evans, J.St.B.T. & Clibbens, J. (in press). Background beliefs and evidence interpretation. Thinking and Reasoning.

Feeney, A., Evans, J.St.B.T. & Venn, S. (2000) The effects of beliefs about the evidence on hypothesis testing. Unpublished manuscript, Department of Psychology, University of Durham.

Gigerenzer, G., Todd, P. & the ABC Research Group (1999). Simple heuristics that make us smart. Oxford: Oxford University Press.

Klayman, J. (1995). Varieties of confirmation bias. In J. Busemeyer, R. Hastie & Medin, D.L. (Eds.), The psychology of learning and motivation, Vol. 32: Decision making from a cognitive perspective (pp 385-419). San Diego: Academic Press.

Klayman, J. & Ha, Y-W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. Psychological Review, 94, 211-228.

Lopez, A. (1995). The diversity principle in the testing of arguments. Memory and Cognition, 23, 374-382.

Mynatt, C.R., Doherty, M.E. & Dragan, W. (1993). Information relevance, working memory and the consideration of alternative. Quarterly Journal of Experimental Psychology, 46A, 759-778.

Oaksford, M. & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. Psychological Review, 101, 608-631.

Oaksford, M., Chater, N. & Larkin, B. (1999). Probabilistic effects in data selection. Thinking and Reasoning, 5, 193-243.

Osherson, D.N., Smith, E.E., Wilkie, O., Lopez, A. & Shafir, E. (1990). Category-based induction. Psychological Review, 97, 185-200.

Popper, K.R. (1959). The logic of scientific discovery. Hutchinson: London.

Spellman, B.A., Lopez, A. & Smith, E.E. (1999). Hypothesis testing: Strategy selection for generalising versus limiting hypotheses. Thinking and Reasoning, 5, 67-91.

Wason, P.C. (1960). On the failure to eliminate hypotheses in a conceptual task. Quarterly Journal of Psychology, 12, 129-140.

Wason, P.C. (1966). Reasoning. In B.M. Foss (Ed.), New horizons in psychology (Vol. 1). Harmondsworth, UK: Penguin Books.

Acknowledgements

This research was funded by grant R000222426 from the Economic and Social Research Council of the United Kingdom.