Title
Probabilities, Utilties and Hypothesis Testing

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 19(0)

Authors
Feeney, Aidan
Evans, Jonathan St.B.T.
Clibbens, John

Publication Date
1997

Peer reviewed

Probabilities, Utilities and Hypothesis Testing

Aidan Feeney, Jonathan St.B.T. Evans and John Clibbens
Department of Psychology
University of Plymouth
Drake Circus
Plymouth
Devon PL4 8AA
United Kingdom
afeeney@plymouth.ac.uk

Abstract

This paper considers the class of hypothesis testing tasks purporting to demonstrate pseudodiagnosticity. It argues that, as has recently been done with other hypothesis testing tasks, pseudodiagnosticity tasks may be re-analysed in terms of people's background beliefs about the probability of their evidential items and the utility of their various test outcomes. A sample analysis of a simplified task is presented along with the results of an experiment which demonstrate that subjects' behaviour corresponds to the prescriptions of the analysis. How the sample analysis might be applied to the standard pseudodiagnosticity task is discussed as are the implications of the results for current accounts of the effects of subjective probability on human hypothesis testing.

Introduction

Traditional, Bayesian, prescriptions about what evidence should be selected to test some set of hypotheses are very straightforward: the selected evidence should allow for an application of Bayes theorem. For example, consider the following scenario presented to subjects by Mynatt, Doherty and Dragan (1993):

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 had no major mechanical problems in the two years she's owned it.

Embedded in this scenario are two hypotheses (one concerning car Xs and the other concerning car Ys) and two features (doing over 25 mpg and having no major mechanical problems in the previous two years). For ease of exposition we will refer to the hypothesis that the subject's sister owns a car X as H_X; to the hypothesis that she owns a car Y as H_Y; to doing over 25 miles per gallon as F_{25mpg}; and to having no major mechanical problems as F_{NMMP}. These hypotheses and features give rise to four pieces of information which are: $P(F_{25mpg}/H_X)$; $P(F_{25mpg}/H_Y)$; $P(F_{NMMP}/H_X)$; and $P(F_{NMMP}/H_Y)$. In order to apply Bayes theorem to Mynatt et al's problem, these pieces of evidence need to be selected in pairs. Thus the subject should select either both pieces of information about the fuel economy of X and Y cars or both pieces of information about their reliability. Doherty, Mynatt, Tweney and Schiavo (1979) using a task structurally analogous to the one above found that subjects tended not to select information in such pairs. Doherty et al termed their subjects' behaviour as pseudodiagnostic.

We agree with the Doherty et al and Mynatt et al analyses of the task as it was used in the 1979, and subsequent (e.g. Kern and Doherty, 1982), studies. However, in recent years there has been an explosion of alternative normative analyses of standard cognitive tasks (e.g. Klayman and Ha, 1987; Anderson, 1991). Many of these analyses rely on subjects' legitimate expectations about these tasks in the light of their background knowledge about the world. Good examples of such analyses are to be found in recent work on Wason's selection task (Oaksford and Chater, 1994; Kirby, 1994; Evans and Over, 1996b) where it is claimed that subjects use their background beliefs about the probability of events in their environment in order to decide which are the best cards to select for the purposes of testing the experimental rule. We believe that such recent insights lead to an alternative normative analysis of certain hypothesis testing tasks.

As stated above, we agree that when faced with a scenario such as the one concerning their sister's car and in the absence of any likelihood information, subjects should choose two pieces of information which, combined, form a likelihood ratio. However, subjects in the Mynatt et al (1993) study exposed to the sister's car scenario also received an initial piece of information concerning $P(F_{25mpg}/H_X)$ and were simply asked to choose one further piece of information from amongst the remaining three in order to help them decide whether their sister's car was a model X or a model Y. Despite this subtle procedural variation, Mynatt et al claimed that the correct normative choice for subjects was $P(F_{25mpg}/H_Y)$. Once again, this was because such a choice allowed subjects to complete the likelihood ratio for F_{25mpg}. We contend that what seems like a simple procedural variation changes the normative analysis of the task which is appropriate. The difference of course lies in whether subjects are given an initial piece of information. If they are, then the task becomes one of selecting the most informative of the remaining pieces of evidence in the light of that initial piece of information. In the absence of an initial piece of information the normatively correct selections will always form a likelihood ratio.

A Sample Analysis

We will illustrate the above claim with a worked example of a simplified version of the pseudodiagnosticity task (a structurally similar version of which has also been used by Doherty, Chadwick, Garavan, Barr and Mynatt, 1996 to investigate the extent to which subjects seek information about the hypothesis about which they already possess some information). This task is simpler because all information is given in non-numerical form and subjects do not receive conditional probabilities. Thus, given the car scenario described earlier, subjects are initially told that model X cars do 25 miles per gallon. The remaining possible pieces of evidence are presented in terms of whether instances of each hypothesis possess each of the features. The structure of this version of the task is presented in Table 1.

Table 1: The structure of a simplified version of the pseudodiagnosticity task

Initial Evidence	Remaining Evidence	Possible Outcomes	Results if Outcome YES	Results if Outcome NO
A. Instances of category X possess initial feature.	B. Whether instances of category Y possess the initial feature.	YES or NO	Uncertainty. Both categories possess the initial feature.	Certainty. The object must be an instance of X.
	C. Whether instances of category X possess the second feature.	YES or NO	Dependent on inference about the diagnosticity of the evidence.	Certainty. The object must be an instance of Y.
	D. Whether instances of category Y possess the second feature.	YES or NO	Uncertainty. Each of the categories possesses at least one feature.	Certainty. The object must be an instance of X.

As may be seen from the structural analysis of the task presented in Table 1 there is no normatively correct Bayesian piece of evidence to be selected. However, because of the structure of the task, there is a best choice - given the assumption that some piece of information will differentiate between the hypotheses. Subjects choosing pieces of evidence B and D can only be certain which of the hypotheses is the case if they obtain NO outcomes. On the other hand, they may be certain that hypothesis X is the case regardless of the outcome arising from a C selection (assuming of course that both model X and model Y cars do not possess both of the features).

The task presented in Table 1, however, may also be analysed in decision theoretic terms. Such an analysis would rest on the Expected Information Gain (Oaksford and Chater, 1994) or Epistemic Utility (Evans and Over, 1996a&b) associated with each card and involves conceptualising the task as one of deciding which piece of information to select. Next, the formula standardly used in decision theory (see Equation 1 where s_i refers to the ith possible outcome of a choice, U_i represents the utility of that outcome, and where i ranges over a finite set of mutually exclusive and exhaustive outcomes) to calculate subjective expected utility of any choice may be adapted for the task.

$$SEU = \sum_i s_i U_i \qquad \text{Equation 1}$$

From Table 1 we can see the approximate utilities for both a YES and a NO outcome for each possible information selection. In all cases a NO outcome allows subjects to differentiate between the hypotheses whereas YES outcomes arising from B and D selection do not reduce uncertainty. The only YES outcome leading to a reduction in subjects' uncertainty is that which follows C selection (and this is dependent on the assumption that some piece of evidence does allow the hypotheses to be differentiated).

The probabilities associated with these various outcomes have not yet been addressed however. In practice, these probabilities will depend on problem content but for the purposes of this worked example consider two features: a rare feature with a probability of 0.3 in the relevant population (for example, having a top speed of over 140 miles per hour in the car problem); and a common feature with a probability of 0.9 (for example, having a radio in the car problem). The probabilities and utilities of each outcome for each of the three remaining pieces of information are given in Table 2. As may be seen from this table the information which any piece of evidence will yield is dependent on the rarity of the feature which it concerns.

Table 2: The probabilities and utilities associated with each possible outcome from the sample task

Initial Feature		B - whether instances of Y possess initial feature		C - whether instances of X possess second feature		D - whether instances of Y possess second feature	
		YES	NO	YES	NO	YES	NO
Common	Prob (D/H)	.9	.1	.9	.1	.9	.1
	Utility	0	m	$0 \leq n \leq m$	m	0	m
Rare	Prob (D/H)	.3	.7	.9	.1	.9	.1
	Utility	0	m	$0 \leq n \leq m$	m	0	m

There are three utility values used in Table 2 - m, n and 0 which are sensitive to the structural analysis of the task presented in Table 1. They also reflect the fact that SEU theory cannot fully capture the logic of that structural analysis. Assuming that there exists some piece of information which differentiates between the hypotheses then there are four possible outcomes which lead to certainty. If the measure of the utility of an outcome which is used happens to be the absolute log likelihood ratio (Equation 2) as is advised by both Evans and Over (1996a&b) and Laming (1996), problems start to arise when one talks of verification or falsification in absolute terms. For example, if the probability of H given a NO outcome to a B, C or D selection is 1 then the

$$Utility = ABS\left[Log\left\{ \frac{\Pr ob(E/H)}{\Pr ob(E/not-H)} \right\} \right] \quad \text{Equation 2}$$

probability of not-H must be 0. This means that the likelihood ratio becomes, very inconveniently, infinity. Although SEU theory's inability to capture deductive certainty may seem very inconvenient at first, it is psychologically plausible. As Evans and Over (1996b) argue, it is unlikely to be the case that no uncertainty remains concerning a set of hypotheses after an observation or set of observations has been carried out. For these reasons, m (the utility assigned to NO outcomes arising from any of the information selections) should be understood as a number tending towards infinity. For the same reasons, assigning a value of 0 to YES outcomes to B and D selections is, strictly speaking, implausible. These outcomes should have utilities close, but not equal to, 0. For ease of exposition, however, they will be assigned utilities of 0.

The only utility value remaining to be discussed is the utility of n assigned to a YES outcome arising from a C selection. Given the assumption that some piece of information will differentiate between the hypotheses, the actual value of n should equal the value of m_. That is, if instances governed by only one of the hypotheses possess both features, a YES outcome arising from a C selection leads to almost complete certainty that hypothesis X is the case. However, in recognition of the fact that such an assumption must be made, a utility of n has been assigned which will decrease towards 0 as the differentiation assumption is considered less safe.

Table 3: Expected information yield for each possible piece of information dependent on the rarity of the initial feature.

Choice	Initial Feature	Expected Information Yield
B - whether instances of Y possess the initial feature	Common	.1m
	Rare	.7m
C - whether instances of X possess the second feature	Common	.9n + .1m
	Rare	.9n + .1m
D - whether instances of Y possess the second feature	Common	.1m
	Rare	.1m

The information to be gained from selecting each piece of information is shown in Table 3. Of most interest is the fact that the expected information yield from a B selection is greatly increased when the feature about which subjects initially receive some information is rare. If the initial feature is common, however, the expected information yield from a C selection will always exceed that of a B selection. In order to conform to this analysis subjects should be more likely to select C when the initial piece of information concerns a common feature than when it concerns a rare feature.

Unfortunately, none of the studies using variants of the pseudodiagnosticity task controlled for the probability of the initial feature. Therefore, we do not know the extent to which subjects' behaviour will conform to the analysis outlined above. The experiment to be presented will test the descriptive accuracy of the analysis of the simplified task presented here whilst the issue of generalising this analysis to the probabilistic version of the task will be addressed in the discussion.

Experiment

The experiment to be described tests the descriptive accuracy of the analysis presented in the pervious section. It has two aims: firstly to see whether the effect of manipulating the rarity of the feature about which subjects initially receive some information is as predicted by the analysis; and secondly to test whether subjects are making the assumption that some piece of information will differentiate between the hypotheses.

Method

Subjects: Of the 96 first year psychology students from the University of Plymouth who participated in this experiment 28 were male and 68 were female. Their mean age was 21. The youngest subject was 18 whilst the oldest was 36.

Materials: each subject received a handout which comprised an instruction sheet and four problems. The instructions given to half of the subjects were as follows:

Accompanying these instructions is a series of four decision problems. Each consists of a description of a situation. The following is an example of the type of situation we have used:

Your friend has just bought a new television. You can't remember whether it's a model X or a model Y but you do remember that it has teletext and a remote control.

Next you will be given a piece of information about the situation. These pieces of information are given in terms of the question you would ask to receive the information and answers which actually tell you what you want to know. For the example situation above you might be told

Question A

whether model X televisions have remote controls

Answer: Yes

Now you know that model X televisions do have remote controls.

*Following each piece of information you will be given a list of the three further questions you could ask with their possible answers. You will be asked to rate the potential usefulness of each question in helping you to decide between the X and Y alternatives present in the description of the situation. If you feel that knowing the answer to a particular question would be extremely helpful in deciding between the two alternatives, you should place a mark at the **"OF GREAT USE"** end of the rating scale. If you feel that knowing the answer to a question would not be helpful in deciding between the two alternatives you should place your mark at the **"OF LITTLE USE"** end of the scale. Remember! The more or less useful that you think a question might be, the closer to the appropriate end of the scale you should place your mark.*

After you have filled in the rating scales you will be asked to select the question which you think would be most helpful in deciding between the X and Y alternatives. You

may feel that you would like to ask more than one question, but please pick only one. We are interested in which question you think would be most useful in helping you make a decision, even though ideally receiving the answer to more than one question might be useful.
It is very important that you read the problems carefully and think about them before filling in the rating scales or picking the question you think would be most useful in deciding between the two alternatives. Take your time and consider your choice before you respond.
If you have any questions at any time, please raise your hand and the experimenter will help you.

The problems used were similar in structure to those used by Doherty et al (1996). In this experiment however, subjects were asked, on a line 100 mm long, to rate the usefulness of each piece of information before making a choice. Below is an example of one of the problems used.

Your sister bought a new car in 1988. You can't remember whether it's a model X or a model Y but you do remember that it has four doors and a radio.
We have already asked the following question for you and have given you the answer:

Question A
whether model X cars bought in 1988 have a radio
Answer: YES
Three additional questions are possible which we have listed below along with their possible answers:

Question B
whether model Y cars bought in 1988 have a radio
Possible Answers: YES or NO
How useful would knowing the answer to this question be in deciding between the alternatives?

OF LITTLE USE　　OF GREAT USE

Question C
whether model X cars bought in 1988 have four doors
Possible Answers: YES or NO
How useful would knowing the answer to this question be in deciding between the alternatives?

OF LITTLE USE　　OF GREAT USE

Question D
whether model Y cars bought in 1988 have four doors
Possible Answers: YES or NO
How useful would knowing the answer to this question be in deciding between the alternatives?

OF LITTLE USE　　OF GREAT USE

Assuming that you could discover the answer to only one of these questions (B, C, or D), which would you ask in order to help you decide which model car your sister drives?

Please circle your choice　B　　C　　D

Half of the subjects received problems such as the one above. The other half were told that their sister's car possessed four doors and a top speed of over 165 mph. This manipulation will be called the Rarity manipulation and has two levels: the Rare level where subjects were given an initial piece of information concerning a rare feature of the class of object to be categorised; and the Common level where subjects were given an initial piece of information concerning a common feature of the target object. The four problem contents, and the features which were used, are shown in Table 4.

Table 4: Problem contents and features (expectations about the features from pre-test are given in the form of percentages) used in the experiment

	Feature Pairs	
Problem Content	Common	Rare
Engineer	earns £25,000 per annum drives a company car	earns over £60,000 p.a. drives a company car
House	has a garden has a garage	has a swimming pool has a garage
Car	has a radio has four doors	top speed of 165 m.p.h. + has four doors
Spanish Villa	costs £150 per week built in the last twenty years	costs £1000 per week built in the last twenty years

These features were independently pre-tested (on 30 subjects for the engineer, house and car problems and on 24 subjects for the spanish villa problem) for their rarity in the problem context. For the engineer problem the common feature was perceived to be possessed by 49% of engineers, whereas the rare feature was perceived to be possessed by 12% of engineers. The corresponding ratings for the other materials are: 81% and 3% for the house problem; 92% and 6% for the car problem; and 51% and 8% for the spanish villa problem.

The initial piece of information which subjects received always concerned the X alternative. In problems with two common features the initial information always concerned the same common feature, whilst in problems with one rare feature the initial piece of information always concerned that rare feature.

One of the aims of this experiment was to test whether subjects make the assumption that some piece of information will differentiate between the hypotheses. Accordingly, it was decided to include a Hint manipulation in the study. This manipulation was intended to control for the possible effects of subjects' pragmatic assumptions about the overall informativeness of the evidence on their pattern of information selections. Accordingly, half of the subjects were told:

N.B. For each of the four problems there is at least one question you can ask which, regardless of the answer to that question, will allow you to be certain in your decision between the alternatives.

A Latin square design was used to control for the order in which B, C, and D appeared (and hence, the order in which their potential usefulness was rated). Four subjects in each condition were presented with one of the six possible orderings of B, C, and D for each problem. The order in which subjects received the problems was randomised.

Results

Two subjects in the no hint/rare condition failed to complete at least one of their four problems. Accordingly, they will not be included in the following analyses.

Evidence selections: The first dependent variable in this experiment was the total number of C (i.e. a further piece of information about instances of category X) choices made by subjects across the four problem contents. Before presenting an analysis of the results on this measure, however, the pattern of information selection across problem contents will be discussed. Across all conditions, the choice frequencies for problem contents were very similar. On the engineer problem, 31% of selections were of B, 53% of C, and 11% of D. On the house problem the equivalent percentages were 34%, 62%, and 4%. On the car problem they were 33%, 61%, and 6%, and on the villa problem they were 39%, 52%, and 9%. As expected, item C was the most highly selected piece of information, followed by item B and then item C.

The numbers of item B, item C, and item D choices were tabulated for each subject across the four problem contents. The mean numbers of item choices, broken down by the Rarity and Hint manipulations, are given in Table 5. A 2x2 between subjects Anova was carried out on the mean number of item C choices in all four between subjects conditions. A significant main effect was found for Rarity ($F(1, 90) = 4.31$, $\underline{p} <. 05$). Neither the main effect of the Hint manipulation ($F(1, 90) = 1.00$, $\underline{p} > .3$), nor the interaction between Hint and Rarity ($F(1, 90) = 2.01$, $\underline{p} > .15$) were found to be significant. These results are as predicted by the normative analysis above.

Table 5: Mean item selections (with standard deviations in bold) broken down by features and instructional manipulation

	COMMON			RARE		
	B	**C**	**D**	**B**	**C**	**D**
HINT	1.00	3.00	0.00	1.46	1.92	0.63
	1.44	**1.44**	-----	**1.28**	**1.61**	**0.77**
NO	1.42	2.25	0.33	1.73	2.05	0.23
HINT	**1.50**	**1.42**	**0.56**	**1.64**	**1.53**	**0.53**
	1.21	2.63	0.17	1.59	1.98	0.44
	1.47	**1.47**	**0.43**	**1.45**	**1.56**	**0.69**

Usefulness ratings: The second dependent measure involved in this experiment was Usefulness Ratings. Subjects' ratings were converted to scores on a 100 unit scale, running from 1 to 100, where each millimetre on the line corresponded to one unit on the scale. The higher a subject's score for a particular piece of information the more useful they rated that piece of information to be. A 2x2x3x4 mixed design Anova was used to analyse the results on this measure. The between subjects factors were Rarity and Hint, whilst the within subjects factors were Item Rated and problem content. The significant main effect for Item Rated ($F(2, 182) = 18.25$, $\underline{p} < .001$) was as expected and follow up Tukey HSDs revealed significant differences between all three means. The mean usefulness rating for item B was 64 ($\underline{SD} = 22$). For item C the mean rating was 71 ($\underline{SD} = 18$), and for item D the mean rating was 57 ($\underline{SD} = 18$). Of most interest in this context was the absence of a significant interaction between Rarity and Item Rated. Although subjects in the Rare conditions were less likely to select item C their ratings of the usefulness of item C did not differ from those of subjects in the Common conditions.

Discussion

The finding of a significant main effect for feature rarity in subjects' evidence selections suggests that their behaviour conforms to the analysis presented in the earlier part of this paper. Subjects given an initial piece of information concerning a common feature were significantly more likely to chose further information relating to the hypothesis about which they already possessed some information than were subjects who received information about a rare feature. According to our decision theoretic analysis, the information to be gained from further evidence concerning the focal hypothesis may be exceeded by the information gained by choosing to discover whether instances governed by the complementary hypothesis also possess the initial feature, but only when the initial feature is rare. The results reported here suggests that subjects are sensitive to feature rarity when selecting evidence with which to test hypotheses.

Although subjects do seem to be sensitive to feature rarity when selecting evidence, the absence of a significant interaction between Item rated and Rarity amongst subjects' Usefulness Ratings suggests that subjects are not aware that rarity affects their choice. The experiment's failure to produce such a significant interaction suggests that Rarity affected subjects' information selections in a manner of which they were unconscious.

Such a suggestion begs the question of how important feature rarity is likely to be in accounting for general patterns of information selection. There seem to be at least two schools of thought on this question. For example, Klayman and Ha (1987) argue that sensitivity to the probabilistic structure of the environment is manifested in a general positive test strategy. This would seem to imply a general, all purpose, heuristic-like strategy. Oaksford and Chater (1994), on the other hand, attribute a great deal of processing about the subjective probability of the events governed by the experimenter's rule to subjects participating in experiments on Wason's selection task. Is our sensitivity to the probabilistic structure of our environment best characterised as *rarity equals informativeness* or do we engage in specific calculations in every situation in which we must select information?

One way of answering this question is to return to the standard version of the pseudodiagnosticity task (as promised in an earlier section of this paper). A re-analysis of the standard task would differ from the sample analysis presented earlier in this paper. The simplicity of the task analysed in this paper was due to the fact that subjects did not receive information in the form of conditional probabilities. On the standard version of the pseudodiagnosticity task where subjects do receive information in the form of likelihoods, the numerical likelihood would interact with the rarity of the feature being quantified in determining the information likely to be gained from any subsequent selection.

If, for example, in the car problem the initial piece of information concerned the percentage of car Xs with a top speed of over 165 mph ($P(F_{165mph}/H_X)$), it would be the likelihood expression which determined the information likely to be gained from selection of evidence concerning $P(F_{165mph}/H_Y)$. So, if the subject was initially told that 95% of car Xs had a top speed of 165 mph then it would be a good idea to select $P(F_{165mph}/H_Y)$. This is because $P(F_{165mph}/H_X)$ at .95 would be much higher than the expected figure. It is unlikely that $P(F_{165mph}/H_Y)$ would be as high and so information about $P(F_{165mph}/H_Y)$ would be likely to discriminate between the hypotheses. However, if $P(F_{165mph}/H_X)$ is approximately equal to the perceived base rate for that feature then it is probably not a good idea to select information concerning $P(F_{165mph}/H_Y)$. This is because it is likely that $P(F_{165mph}/H_Y)$ will also equal the base rate and the evidence will not discriminate between the hypotheses.

The extent to which subjects behave in accordance with the prescriptions of such an analysis would be indicative of the degree to which they are sensitive to rarity information. For example, the finding that rarity and likelihood acted independently in affecting subjects' behaviour would suggest the operation of a *rarity equals informativeness* heuristic. On the other hand, if these factors were found to interact so that subjects' behaviour was in close agreement with that of the informal model sketched above, then rarity information could be inferred to have affected subjects' behaviour in a relatively complex manner. The first two of the present authors have recently started to run a set of studies testing subjects' adherence to the principles of the analysis just presented. Initial results suggest that although rarity is an important factor in subjects' pattern of information selection, it operates independently of likelihood. It would seem, therefore, that although *rarity equals informativeness* may be a general purpose heuristic employed by subjects in hypothesis testing, the amount of processing involving specific subjective probabilities which subjects do in such situations may be highly constrained.

Another interesting research question concerns the values assigned by subjects to the utility parameters in the model. These values are likely to depend on subjects' representation of the task. For example, the value assigned to the n-parameter (the utility assigned to the discovery that instances of the focal category possess the second evidential feature) will be high only when the subject infers that some evidence will discriminate between the hypotheses. This inference is more likely to be made by subjects who fail to consider the potential impact of evidence relevant to the complementary hypothesis. Subjects in the present experiment appear to be making this inference but we would predict that any manipulation which caused subjects to consider the complementary hypothesis would decrease the value which they assigned to n and therefore, reduce their tendency to select further information relevant to the focal hypothesis. In this sense, our model is one of constrained rationality and is in accordance with recent proposals by Johnson-Laird and Byrne (1993) and Evans and Over (1996b) who claim that, in many cases, "irrational" inferences and hypothesis testing performance are caused by subjects' failure to consider all logical possibilities or the potential impact of all the available evidence.

Acknowledgements

This work was supported by a postgraduate studentship to the first author from the Faculty of Human Science at the University of Plymouth.

References

Anderson, J.R. (1991). Is human cognition adaptive? *Behavioural and Brain Sciences*, 14, 457-471.

Doherty, M.E. et al. (1979). Pseudodiagnosticity. *Acta Psychologica*, 43, 111-121.

Doherty, M.E. et al. (1996). On people's understanding of the diagnostic implications of probabilistic data. *Memory and Cognition*, 24, 644-654.

Evans, J.St.B.T. and Over, D.E. (1996a). Rationality in the selection task: Epistemic utility versus uncertainty reduction. *Psychological Review*, 103, 356-363.

Evans, J.St.B.T. and Over, D.E. (1996b). *Reasoning and rationality*. Hove, U.K.: Lawrence Erlbaum Associates Ltd.

Johnson-Laird, P.N. and Byrne, R.M.J. (1993). Models and deductive rationality. In K.I. Manktelow and D.E. Over (Eds.), *Models of rationality*. London: Routledge.

Kern, L. and Doherty, M.E.(1982). "Pseudodiagnosticity" in an idealized medical problem-solving environment. *Journal of Medical Education*, 57, 100-104.

Kirby, K.N. (1994). Probabilities and utilities of fictional outcomes in Wason's four-card selection task. *Cognition*, 51, 1-28.

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

Laming, D. (1996). On the analysis of irrational data selection: A critique of Oaksford and Chater. *Psychological Review*, 103, 364-373.

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

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