Normative and Information Processing Accounts of Medical Diagnosis

Peter Yule (P.Yule@psyc.bbk.ac.uk)
Department of Psychology, Birkbeck College, University of London, Malet St.,
London, WC1E 7HX

Richard Cooper (R.Cooper@psyc.bbk.ac.uk)
Department of Psychology, Birkbeck College, University of London, Malet St.,
London, WC1E 7HX

John Fox (jf@acl.icnet.uk)
Advanced Computation Laboratory, Imperial Cancer Research Fund, Lincoln's Inn Fields,
London, WC2A 3PX

Abstract

The field of Judgement and Decision Making has for some time been dominated by normative theories which attempt to explain behaviour in mathematical terms. We argue that such approaches provide little insight into the cognitive *processes* which govern human decision making. The dominance of normative theories cannot be accounted for by the intractability of processing models. In support of this view, we present a processing account of performance on a simulated medical diagnosis task. The performance of the model, which includes learning, is compared with that of a normative (Bayesian) model, and with subject performance on the task. Although there are some caveats, the processing model is found to provide a more adequate account of subject performance than the Bayesian model.

Introduction

A tension arises in many areas of cognitive psychology between mathematical accounts and processing accounts of behaviour. Mathematical accounts attempt to develop equations with which behaviour can be described or predicted. This typically results in normative theories. Interest then focusses on systematic departures from normative behaviour. Processing accounts are less concerned with numerical relationships between stimuli and responses. Instead, these accounts focus on the underlying sequence of informational states that a cognitive agent progresses through in the processing of stimuli leading up to the generation of a response. The tension is exemplified by the field of Judgement and Decision Making (JDM), where the dominant approach has, for several decades, been a normative (i.e. mathematical) one.

According to the normative view, human decision making under uncertainty can be described in probabilistic terms. Given a set of options, each with several possible outcomes, the option chosen is that which maximises expected utility, where the expected utility of an option is the sum of the subjective value of each possible outcome multiplied by the subjective probability of each outcome given that the option is chosen. The cost of each option may also be included in this calculation. (See, for example, Lindley, 1985.) Attention within the normative JDM community has focussed on the investigation of heuristics which influence the relationship between objective and subjective probability and between objective and subjective value. (See, for example, Tversky & Kahneman, 1974.)

Despite the arguable successes of the normative approach to JDM, such approaches say little about the cognitive processes underlying decision making. There is now, however, a growing body of research which addresses these processes. Fox (1980), for example, studied a simplified medical diagnosis task, in which each subject played the role of a doctor attempting to diagnose a series of patients. Within the task (described in more detail below) symptoms were probabilistically associated with diseases, such that, for example, headache occured in 75% of cases of meningitis. Subjects were given several blocks of trials in which to learn the task. On a final block of trials three dependent measures — diagnostic accuracy, number of symptoms queried, and the order in which symptoms were queried — were collected.

The relevant normative model in the case of diagnosis tasks is based on the application of Bayes' theorem. In Fox's task this allows, for each disease, the probability that a patient has that disease given the patient's symptoms and the probability of each symptom given each disease, to be calculated. Fox (1980) developed mathematical (i.e. Bayesian) and processing accounts of subjects' performance in the task. The Bayesian model yielded a reasonable fit to all three measures of subject performance, but this fit was at least matched by the symbolic model. Fox & Cooper (1997) replicated these results with a first-order, domain independent, version of the symbolic model.

In further work, Cooper & Fox (1997) reported learning data for the task. This data is derived from subjects' performance over a series of blocks of trials. Subjects begin the task in a naive state, performing at chance levels (which, with five possible diseases, corresponds to a diagnostic accuracy of 20%). After three blocks many subjects achieve a diagnostic accuracy of over 80%. Two conditions, differing in the average number of symptoms associated with each disease, were explored in the learning task. Cooper & Fox (1997) also extended the first order symbolic model of Fox & Cooper (1997) to provide an account of learning. Only two dependent mea-

Table 1: Conditional probabilities of symptoms given diseases in each matrix condition.

Matrix		Mesiopathy	Ritengitis	Katalgia	Bonanoma
Dense	Diarrhoea	1.00	0.50	1.00	0.00
	Fever	0.00	1.00	1.00	1.00
	Headache	0.75	0.00	1.00	0.75
	Paralysis	0.75	0.00	0.75	1.00
	Vomiting	1.00	0.50	0.00	1.00
Sparse	Diarrhoea	0.00	0.50	0.00	1.00
	Fever	1.00	0.00	0.00	0.00
	Headache	0.25	1.00	0.00	0.25
	Paralysis	0.25	1.00	0.25	0.00
	Vomiting	0.00	0.50	1.00	0.00

sures were collected in the learning task (diagnostic accuracy and number of symptoms queried), but the extended model exhibited the observed qualitative differences in these measures, both within and between the conditions.

This paper continues the research programme by 1) reporting new data replicating and extending that reported by Cooper & Fox (1997); and 2) extending the Bayesian model of the task to include learning and question ordering strategies. The replication reported here includes the third dependent measure, the order in which symptoms were queried, not previously recorded for the learning version of the task. These data, for the final block of trials, are used in a *one-ply analysis*, where we compare subjects' initial behaviour given a presenting symptom — whether they question any of the remaining symptoms or offer a diagnosis — with predictions derived from both models' initial behaviour. Our results again support the view that the symbolic model captures human performance at least as well as the Bayesian model.

The rest of the paper begins by introducing the task in some detail. We then describe the relevant features of each of the computational models, before reporting empirical findings. We conclude by evaluating each of the models in the light of our empirical findings.

The Task

Within the simulated medical diagnosis task subjects are required to diagnose a number of "patients". There is a fixed set of symptoms and diseases from which each patient might be suffering. Five symptoms and four diseases are used in the experiments reported here. Each patient is suffering from one and only one disease. This disease manifests itself in the patient's symptoms. Each patient has at least one symptom, the presenting symptom. In addition patients may have any or all of the four remaining symptoms.

Each trial begins with the subject being told the current patient's presenting symptom. The subject may then query the presence or absence of any of the remaining symptoms, and can offer a diagnosis at any point. When a diagnosis is given, the subject is informed of its correctness, and, if incorrect, of the disease the patient was actually suffering from. It is this feedback that allows subjects to learn the task.

The task was designed to be relatively naturalistic, whilst maintaining a high level of experimental control. Of particular importance are the number and variety of measures of performance which the task makes available. At a purely numeric level, each block of trials provides measures of diagnostic accuracy and mean number of questions asked before diagnosis. Importantly, the task also provides information concerning sequential processing by subjects in the form of question ordering preferences.

The task is also open to a variety of manipulations. As noted above, Cooper & Fox (1997) reported data from two conditions differing in the average number of symptoms associated with each disease. These are further explored here. In the "sparse" condition, each disease is characterised by relatively few symptoms. In the "dense" condition each disease has the reverse symptom pattern to that of the sparse condition, so each disease has relatively many symptoms.

The presence of symptoms is probabilistic, with symptom patterns being generated from tables of conditional probabilities of symptoms given diseases, and presenting symptoms are selected according to a function weighted by the symptoms' conditional probabilities given diseases. Table 1 shows the conditional probabilities of symptoms given each disease in each condition. In one sense, the dense and sparse conditions are informationally identical: given that symptoms are either present or absent, presence of a symptom in one condition is informationally equivalent to absence in the other condition and *vice versa*. In another sense the conditions differ significantly: patients always present with a symptom (rather than without a symptom). In the dense condition more diseases are associated, on average, with each positive symptom than in the sparse condition. Hence, the initial presenting symptom carries less information in the dense condition than in the sparse condition.

Computational Models

Two models of performance of the task — one Bayesian and one Symbolic — have been developed within the COGENT modelling environment (Cooper & Fox, in press). CO-

GENT provides a number of facilities which simplify the development and specification of cognitive models, including a graphical interface to a rule based modelling language, and tools for Monte Carlo simulation of multiple subjects.

The Basic Model

Figure 1 shows the functional modules (as they appear in the COGENT specification) of both the Bayesian and Symbolic diagnosis models with learning.

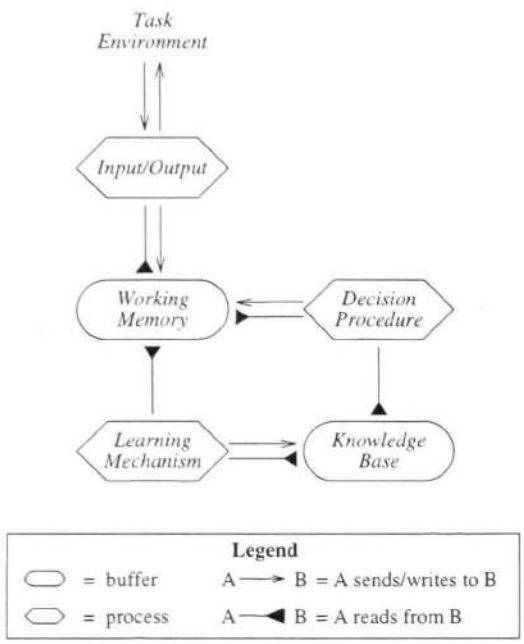

Figure 1: Box/Arrow diagram of the model with learning, from Fox & Cooper (1997).

The function of each of the boxes is as follows. *Task Environment* (which is not shown in detail) generates subject data, presents it to the rest of the model, answers queries concerning the presence/absence of symptoms, and records all protocols. It is not part of the cognitive model but is implemented within COGENT so as to automate the data presentation and analysis. *Input/Output* models the subject's perceptual/articulatory processes. Messages from *Task Environment* trigger additions to *Working Memory* (e.g., adding information about the presence of a symptom), and the existence of appropriate elements in *Working Memory* trigger generation of articulatory output (e.g., a query about a symptom). *Working Memory* is a passive data store in which information about the current case is stored and manipulated. There is no limit on, or decay of, the information stored here, and in both models information is retrieved in the same order as it is added, that is, access is First-In/First-Out.

Decision Procedure is a set of inference rules which modify *Working Memory*, implementing the basic diagnostic strategy, which differs for the Bayesian and Symbolic models. The details of the Symbolic model have already been published in Cooper & Fox (1997), but to summarise briefly, receipt of a `told(Symptom, Value)` message in *Working Memory* triggers a rule in *Decision Procedure* which augments *Working Memory* with the set of diseases which are suggested by the presenting symptom. The presence of suspected diseases in *Working Memory* then prompts recall of their associated symptoms (through a second *Decision Procedure* rule). One of these symptoms is then selected as a query, and sent to *Input/Output*. Diagnosis depends on a *Decision Procedure* rule which matches the symptom pattern with characteristic symptom patterns stored in *Knowledge Base*; symptoms are queried sequentially until the diagnosis rule succeeds in matching a stored pattern.

In the Bayesian model, by contrast, receipt of a `told(Symptom, Value)` message in *Working Memory* triggers a rule in *Decision Procedure* which calculates the conditional probabilities of all the possible diseases given the symptom information currently available, and adds these to *Working Memory*. The presence of these conditional probabilities is used in the next processing cycle to either make a diagnosis, in the case that one of the diseases already has a conditional probability greater than the threshold value, or otherwise, to select another symptom to query. This is accomplished by a rule in *Decision Procedure* which calculates the overall informativeness of each remaining symptom query, adding these to *Working Memory*. The symptom with the greatest informativeness value is then selected as the next symptom query, and an appropriate message is sent to *Input/Output*. The Bayesian model therefore uses expected utility theory (where utility corresponds to informativeness) supplemented with Bayes' theorem to select each question.

In both models, the rules in *Decision Procedure* use task knowledge stored in *Knowledge Base*. Both models learn this knowledge as the task progresses.

Learning

As detailed in Cooper & Fox (1997), the Symbolic model learns by storing a variety of types of information in *Knowledge Base*; the three types of information stored are:

1. clauses detailing which diseases are suggested by which symptoms;
2. clauses specifying whether symptoms and diseases are positively or negatively associated; and
3. clauses specifying typical patterns of symptoms for each disease.

The Bayesian learning model is even simpler: its learning procedure just counts frequencies, both of symptoms and of symptoms given diseases. These are used by the diagnosis procedure to estimate conditional probabilities.

In both models, the *Knowledge Base* has Last-In/First-Out access, resulting in an overall recency effect. This does not affect the operation of the Bayesian model, but does affect the Symbolic model's questioning strategy.

Table 2: Mean % diagnostic accuracy (Symbolic model).

Matrix	Blk 1	Blk 2	Blk 3	Blk 4	Mean
dense	57.50	89.00	94.50	99.50	85.13
sparse	63.00	79.00	78.00	75.00	73.75
Mean	60.25	84.00	86.25	87.25	79.44

Table 3: Mean symptoms queried (Symbolic model).

Matrix	Blk 1	Blk 2	Blk 3	Blk 4	Mean
dense	3.96	3.88	3.83	3.87	3.89
sparse	2.47	0.56	0.42	0.53	1.00
Mean	3.22	2.22	2.13	2.20	2.44

Table 4: Mean % diagnostic accuracy (Bayesian model).

Matrix	Blk 1	Blk 2	Blk 3	Blk 4	Mean
dense	60.50	77.50	82.00	82.00	75.50
sparse	71.50	83.50	85.50	90.50	82.75
Mean	66.00	80.50	83.75	86.25	79.13

Table 5: Mean symptoms queried (Bayesian model).

Matrix	Blk 1	Blk 2	Blk 3	Blk 4	Mean
dense	1.71	1.33	1.46	1.39	1.47
sparse	1.21	0.90	0.99	1.15	1.06
Mean	1.46	1.12	1.23	1.27	1.27

Modelling Experiments

Each learning model was run 10 times in each matrix condition for 4 blocks of 20 trials each, and percentage accuracy and mean number of symptoms queried in each block were recorded. Also, the first question asked in each trial of the final block (or diagnosis if no questions were asked) was recorded. Parameter settings for all runs were identical, so the only source of variation in their behaviour was the particular stimulus set used on each run.

For the Bayesian model, a diagnosis threshold value of 0.60 was chosen. This gives broadly similar diagnostic accuracy in both models.

Symbolic model Table 2 shows the mean percentage diagnostic accuracy scores in each block and matrix condition for the Symbolic model. There is a highly significant effect of block ($F(3,54) = 38.31$, $p < 0.001$), such that overall diagnostic accuracy increases across the blocks, and there is a significant effect of matrix ($F(1,18) = 42.61$, $p < 0.001$), such that performance is higher in the dense than in the sparse condition. Also, they interact ($F(3,54) = 9.36$, $p < 0.001$); whereas the dense learning curve continues to rise across the blocks, the sparse one flattens off at a lower level.

Table 3 shows mean numbers of symptoms queried by the Symbolic model, in each block and matrix condition. There are highly significant effects of both block ($F(3,54) = 94.85$, $p < 0.001$) and matrix ($F(1,18) = 1161.34$, $p < 0.001$), and an interaction between them ($F(3,54) = 76.03$, $p < 0.001$). In the dense condition, the number of symptom queries remains close to four in each block, whereas in the sparse condition the number of queries drops to an average of less than one from the second block onwards.

Bayesian model Table 4 shows the mean percentage diagnostic accuracy scores in each block and matrix condition for the Bayesian model. There is a highly significant effect of block ($F(3,54) = 29.14$, $p < 0.001$), with accuracy increasing from block 1 to 4, and a significant effect of matrix ($F(1,18) = 8.55$, $p < 0.01$), such that diagnostic accuracy is higher in the sparse condition, unlike the Symbolic model. This time there is no interaction ($F(3,54) = 0.92$).

Table 5 shows mean numbers of symptoms queried by the Bayesian model, in each block and matrix condition. There is a significant effect of Block ($F(3,54) = 18.00$, $p < 0.001$), such that the number of symptoms drops from block 1 to 2, but then gently rises again. There is also a significant effect of matrix ($F(1,18) = 40.13$, $p < 0.001$), such that more symptoms are queried in the dense condition, and a barely significant interaction ($F(3,54) = 2.94$, $p < 0.05$).

One-ply Predictions

The one-ply predictions of each model were generated by aggregating initial selection behaviour, on each trial in the fourth block, across all simulated subjects in each matrix condition. This resulted in a set of tables of percentages of occasions when each symptom, or a diagnosis, was selected, for each presenting symptom. For each model, the two cells with the highest values in each row were then selected as the predictions. When there was a second-place tie, all three cells were included as predictions; this occurred once for each model. These predictions are annotated on Table 8 below.

Experiment

Method

Subjects 38 second year psychology students from Birkbeck College took part, 18 in the dense condition and 20 in the sparse condition.

Design Each S was assigned to either the Dense or Sparse matrix condition, and performed four blocks of 20 trials, each of which comprised 5 trials with each disease, in random sequence. All diagnoses and symptom queries were recorded.

Software The task was computer-based, mouse driven and administered by a client-server system on the departmental intranet, using a network of 486 PCs. The client portion, written in JavaScript for Netscape Navigator 4, randomised trials within blocks, presented stimuli and collected responses. The server assigned subjects to dense and sparse matrix conditions, and collated data.

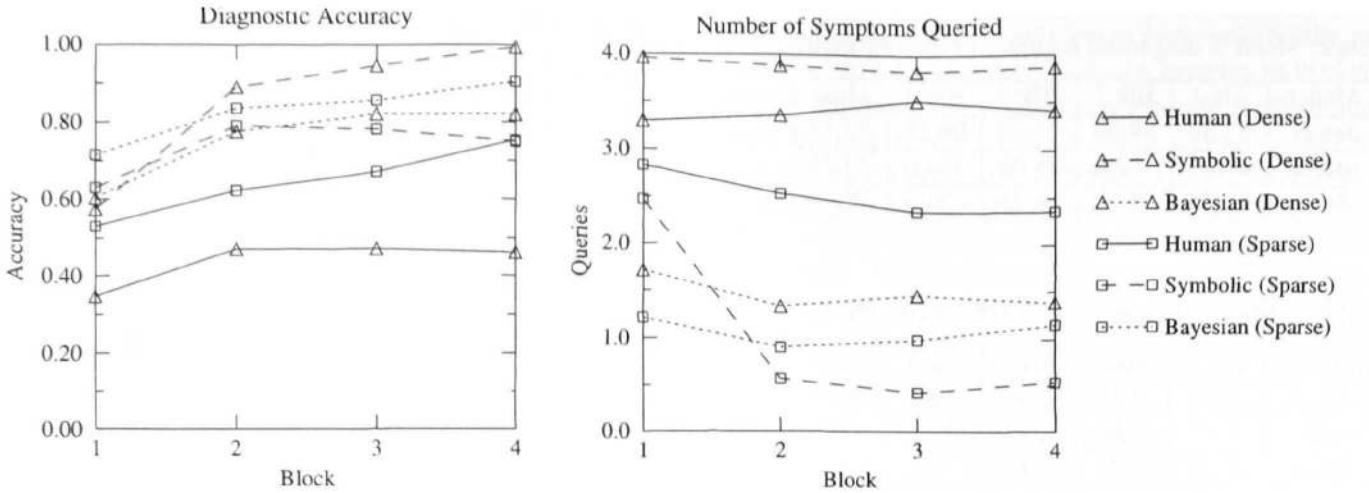

Figure 2: Graphs of diagnostic accuracy (left) and number of symptom queries (right) for human subjects and both models.

Table 6: Mean % diagnostic accuracy (Human data).

Matrix	N	Blk 1	Blk 2	Blk 3	Blk 4	Mean
dense	18	34.70	46.95	46.95	46.10	43.68
sparse	20	53.00	62.25	67.00	75.25	64.38
Mean	38	44.34	55.00	57.50	61.44	54.57

Table 7: Mean symptoms queried (Human data).

Matrix	N	Blk 1	Blk 2	Blk 3	Blk 4	Mean
dense	18	3.29	3.35	3.52	3.41	3.39
sparse	20	2.83	2.53	2.36	2.36	2.52
Mean	38	3.05	2.92	2.91	2.86	2.93

The client system[1] was launched by clicking on a button at the foot of a web page of instructions, which opened a new window. On each trial the program displayed a series of boxes labelled with symptom names, running across the top half of the window, and a series of boxes labelled with disease names running across the bottom. The order in which both symptom and disease boxes were presented onscreen was randomised on each trial.

The symptom boxes could be clicked by the subject, when they would change to reveal whether the symptom was present or absent in the imaginary patient. At the beginning of each trial, one of the symptom boxes was already in this changed state, giving the subject information about the patient's presenting symptom. When the disease boxes were clicked, a new box appeared in the centre of the screen stating whether the diagnosis was correct or not. If the diagnosis was incorrect, this box also gave feedback about the correct disease, allowing the subject to learn. Clicking on this box started the next trial.

At the end of each block, the program presented a score giving the number of correct diagnoses out of 20, saved the accumulated data to the server, and gave the subject the opportunity to rest briefly.

Instructions The launch page of the experimental client system described the screen layout and the block structure of the experiment. Subjects were also verbally instructed to attempt to diagnose efficiently, that is, to minimise the number of symptom queries they made, provided this did not compromise their diagnostic accuracy.

Results

Learning results Table 6 and Figure 2 (left) show mean percentage diagnostic accuracy for human subjects. Human diagnostic accuracy is substantially lower throughout than that achieved by either model. However, there are highly significant effects of both block ($F(3,108) = 16.27$, $p < 0.0001$) and matrix ($F(1,36) = 12.42$, $p < 0.0012$), such that subjects in the sparse condition performed substantially better than those in the dense condition, like the Bayesian but unlike the Symbolic model. Also, there is a barely significant interaction ($F(3,108) = 2.77$, $p < 0.0452$), due to the flattening of the learning curve in the dense condition.

Table 7 and Figure 2 (right) show mean numbers of symptoms queried by human subjects in each block and matrix condition. Human subjects query substantially more symptoms than do either of the models. There is a modestly significant effect of matrix ($F(1,36) = 4.83$, $p < 0.0345$), such that more symptoms are queried in dense than in sparse, but no overall effect of block ($F(3,108) = 1.17$). However, there is a highly significant interaction ($F(3,108) = 5.23$, *Greenhouse-Geisser* $p < 0.0045$), due to the reduction in symptom queries across blocks in the sparse but not in the dense condition; instead the number of queries in the dense condition *rises* from block 1 to block 3. Aside from this increase, this pattern of results is more similar to the Symbolic model than the Bayesian model, as the number of queries in the dense condition remains effectively at ceiling whereas that in the sparse condition drops across the blocks.

One-ply analysis Table 8 shows the human one-ply results, the percentages of each possible initial selection behaviour given each presenting symptom, aggregated across subjects, for both dense and sparse matrix conditions. The predictions of Symbolic and Bayesian models are indicated by superscript B and S respectively.

We evaluate the relative performance of the models by scoring how often each model predicts the top two values in

[1]For a demonstration of the client system, see http://www.psyc.bbk.ac.uk/staff/pgy/experiments/jdm2a/demo/

Table 8: One-ply analysis for Block 4 (Human data). Superscript B and S indicate the respective model predictions.

Matrix	Presenting symptom	N	First query					
			diagnosis	diarrhoea	fever	headache	paralysis	vomiting
dense	diarrhoea	54	1.5%		20.0%^{SB}	26.2%^{SB}	23.1%	29.2%
	fever	123	6.4%	16.7%^{S}		20.6%^{B}	26.2%^{S}	30.2%^{SB}
	headache	64	4.2%	47.9%^{S}	18.8%^{B}	-	12.5%	16.7%^{SB}
	paralysis	48	14.0%	26.3%^{S}	12.3%^{B}	12.3%		35.1%^{SB}
	vomiting	71	3.1%	26.6%^{B}	20.3%^{SB}	35.9%^{SB}	14.1%	-
sparse	diarrhoea	107	34.7%^{SB}		14.9%^{B}	19.8%^{S}	15.8%	14.9%
	fever	90	39.3%^{SB}	5.6%		28.1%	14.6%	12.4%
	headache	46	14.3%^{SB}	30.4%^{S}	25.0%^{B}	-	8.9%	21.4%
	paralysis	58	30.4%	19.6%	4.4%^{B}	15.2%^{SB}		30.4%^{S}
	vomiting	99	25.9%^{S}	14.8%	19.4%^{B}	20.4%^{B}	19.4%^{S}	

each row; each model gets one point for each predicted cell containing one of the two highest values in the row. According to this criterion, the Bayesian model scores 6 points in the dense condition and 4 points in the sparse condition, totalling 10/20, whereas the Symbolic model scores 7 in the dense condition and 6 in the sparse condition, a total of 13/20. So the Symbolic model performs better on this measure than does the Bayesian model.

Discussion

We have presented subject data and two computational models of performance (including learning) on a simulated medical diagnosis task. Each model captures some aspects of the subject data, but although the Bayesian model predicts the greater diagnostic accuracy observed in the sparse condition, the symbolic model arguably gives a better fit with the number of symptom queries, and clearly outperforms the Bayesian model in fitting the one-ply data, so we consider it a better fit overall.

Although the increase in diagnostic accuracy across the blocks by subjects was similar to that predicted by both models, overall diagnostic accuracy performance was poorer than that reported by Fox (1980) and Cooper & Fox (1997), and that predicted by both models, despite the fact that the task was easier than that used by Cooper & Fox. This may be attributable to differences between the populations tested: the experiments were separated by almost twenty years, with medical students in the first case and psychology students in the present case.

Subjects achieved higher levels of diagnostic accuracy in the sparse condition, unlike the Symbolic model. This could be due to the greater memory load in the dense condition, by which the models, being without capacity limitations or decay, are unaffected. We plan to investigate the effects of imposing memory limitations on the models in future.

The fact that the Symbolic model, with a FIFO Knowledge base and LIFO Working Memory, accounted reasonably well for the one-ply data, corroborates the view that recency in the accessibility of symptom information is a major determinant of subjects' questioning behaviour (Fox, 1980; Fox & Cooper, 1997). The present finding extends Fox's earlier results to the case of a learning model, and to a new materials set including the dense/sparse manipulation. In this model the recency effect operates in the long term, that is for periods longer than a single trial, and so recently learned information in the Knowledge Base is the first to be retrieved, and to appear in symptom queries, provided it is relevant. Other, unreported attempts to fit the present data with different buffer access settings were less successful, as expected. We aim to explore this issue further, and to develop other variants of the Symbolic learning model with alternative questioning strategies, in the hope of accounting more fully for the subject data.

References

Cooper, R., & Fox, J. (1997). Learning to make decisions under uncertainty: The contribution of qualitative reasoning. In Shafto, M.G., & Langley, P., (eds.) *Proceedings of the 19th Annual Conference of the Cognitive Science Society.* pp 125-130. Lawrence Erlbaum Associates. Hillsdale, NJ.

Cooper, R., & Fox, J. (in press). COGENT: A visual design environment for cognitive modeling. To appear in *Behavior Research Methods, Instruments, and Computers.*

Fox, J. (1980). Making decisions under the influence of memory. *Psychological Review, 87,* 190–211.

Fox, J., & Cooper, R. (1997). Cognitive processing and knowledge representation in decision making under uncertainty. In Scholz, R. W., & Zimmer, A. C., (eds.), *Qualitative Theories of Decision Making.* pp. 83–106. Pabst, Lengerich, Germany.

Lindley, D. V. (1985). *Making Decisions.* 2nd edition. John Wiley, Chichester, England.

Tversky, A., & Kahneman, D. (1974). Judgement under uncertainty: Heuristics and biases. *Science, 185,* 1124–1131.