BELIEF IN A DIAGNOSTIC HYPOTHESIS AND FEATURE DETECTION

THE EFFECTS OF MANIPULATING THE DEGREE OF BELIEF IN A DIAGNOSTIC HYPOTHESIS ON FEATURE DETECTION

By

VICKI LEBLANC, B.PS.

A Thesis

Submitted to the School of Graduate Studies

in Partial Fulfillment of the Requirements

for the Degree

Master of Science

McMaster University

MASTER OF SCIENCE (1996)
(Psychology)

McMaster University
Hamilton, Ontario

TITLE: The Effects of Manipulating the Degree of Belief in a Diagnostic Hypothesis on Feature Detection.

AUTHOR: Vicki LeBlanc. B. Ps (Université de Moncton)

SUPERVISOR: Professor Lee R. Brooks

NUMBER OF PAGES: vi, 61

Abstract

In Experiment 1, the degree of belief in a focal hypothesis was manipulated using priming as well as the principle of unpacking of Tversky and Koehler (1994). The effects of these manipulations on feature detection was measured. It was found that regardless of the degree of belief in the focal hypothesis, novice diagnosticians who have it in mind will call more of its features than those who do not have it in mind. It is believed that this is due to the fact that having a diagnosis in mind seems to focus the attention of diagnosticians to the relevant features. Also, our manipulation of suggesting alternatives to the diagnosticians did not have the effect of decreasing the diagnosticians' belief in the focal hypothesis, contrary to what is predicted by Tversky and Koehler's unpacking principle (1994). The results from Experiment 1 suggest, and those from Experiment 2 confirm the hypothesis that in order to decrease the degree of belief in the focal hypothesis when it is presented with alternatives, the alternatives must be plausible. If the focal hypothesis is extremely dominant over the alternatives, a reversal of the unpacking principle will occur.

Acknowledgements

I would like to thank my supervisors, Dr. Lee Brooks and Dr. Geoffrey Norman for not only providing me with a great project, but for also giving me the best supervision that any student could ask for. They came so close to convincing me to stay in this field of study. I would also like to thank my committee member, Dr. Bruce Milliken, for providing valuable input during the writing of this thesis.

I would also like to thank the members of my family who have both inspired me and supported me throughout my life. To my parents who, having been through this all, were able to understand, guide and support me through the hard times (as well as rejoice through the good times). To my sister Anik and my brother Stéphane, who have shown me that to take risks and to go after what you want leads to a better life.

I would also like to give thanks to my fellow graduate students at McMaster who in turn, have provided me with words of wisdom, have listened to my woes, and have plainly been fun to be around. Thank you to Karmen Bleile, Chris Horn, Stephanie Hevenor and Dr. Jim Debner. A special thank you goes to the junior god, Tim Wood.

Last but not least, I would like to thank my husband, Denis Daigle, for being there (whether he wanted to be or not) throughout the good, the bad and the evil times.

Table of contents

List of Figures and Tables

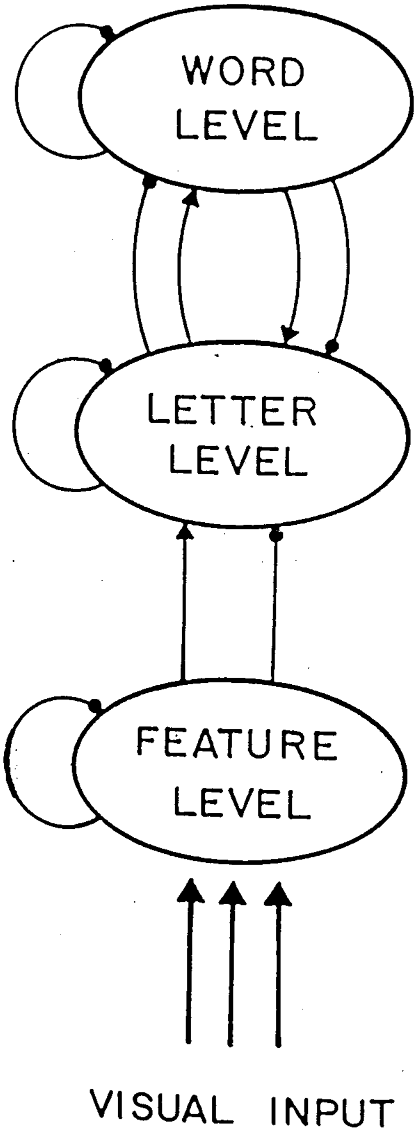
When physicians make medical diagnoses regarding patients, much of the information that they use to come to a diagnostic conclusion is visual. Physical signs, such as rashes, swelling, and color of skin, provide an important source of information as to the nature of a patient's condition. It is based on the observation of these features that clinical tests are ordered, and that potential treatments are considered. Given the observed signs, diagnosticians can come up with several possible diagnostic alternatives. To eliminate some of these alternatives, physicians can use information such as the prevalence of a given disease as well as the sensitivity and the specificity of the features to the disease. The prevalent view in medicine is that this background information should be combined using a Bayesian approach. When applied to the field of medicine, Bayes' theorem states that the degree of belief in a diagnosis given specific features should be revised in light of information regarding (the prevalence of the disease weighted by the likelihood of the disease given the features) relative to (the likelihood of alternative diagnoses given the features weighted by the prior likelihood of the diseases). This type of reasoning is regarded as the ideal procedure to undertake when estimating the degree of belief in a diagnosis, and is often used to evaluate actual behavior of diagnosticians (Sackett, Haynes, & Tugwell, 1985). The Bayesian approach has been implemented in computerized models of medical diagnosis, such as Iliad and Meditel (Berner et al., 1994). Using one of these programs, diagnosticians type in the observed features. The program then generates a list of the most plausible diagnoses on the basis of the listed features. In addition, it indicates which clinical tests would provide additional features that would most efficiently distinguish between the suggested diagnoses.

This Bayesian approach, as well as the computerized diagnostic programs that follow from it, have their strengths in the fact that they emphasize the importance of baseline probabilities, which are often neglected (Griffin & Tversky, 1992). The programs provide diagnosticians with additional diagnoses about which they can gather information regarding a patient's condition. The belief seems to be that programs such as Iliad and Meditel will eliminate diagnostic variability due to availability of diagnoses or ignorance of a relationship between a given feature and a diagnosis. What this approach and the programs fail to do is address the issue of the features that are or are not reported from the data already accumulated. The assumption seems to be that **seeing** a feature that is present is unaffected by factors that could influence the strength of belief in a given diagnosis. By exclusive emphasis on programs that use lists of features as their input, diagnosticians seem to be suggesting that the problem in inaccurate diagnosis resides in the weighting and the combination of the provided evidence, rather than in the extraction of these features.

There is, however, evidence from the field of psychology to suggest that variations in the context can greatly influence the probability of detecting a feature already present. The most commonly known example of this is the Word Superiority Effect (WSE). First demonstrated by Reicher (1969), WSE is defined by the fact that letters are more easily detected in words and orthographically regular pseudowords than when presented alone or in unrelated letter strings (see Whittlesea & Brooks, 1988). A commonly accepted explanation of this phenomenon is found in McClelland and Rumelhart's (1981) interactive activation model of perception. The basic assumption of this model is that, in addition to receiving sensory information, we bring in knowledge about the general properties of objects whenever we are perceiving. In this model, there is a visual feature level, a letter level, a word level, as well as higher levels of processing. These higher levels provide top-down information to the word level regarding what is expected. All of these levels of processing are believed to function in parallel, both spatially and between the levels. That is, the system can process several letters at the same time and it can operate at several levels at the same time. The most important element of this model is that the processing is interactive; the incoming "bottom-up" feature information interacts with the "top-down" information, what we know about words, to determine what we are seeing.

The model functions as follows (see figure 1). At each of the levels mentioned above, there are nodes which represent specific units. Each of these nodes has a different baseline level of activation which is determined by the frequency of the unit in the language. Sensory information comes up from the perceived objects. If enough information comes in consistent with the presence of a feature, then the node for that feature will become activated. In turn, this node has both excitatory and inhibitory connections to other nodes on the same level and on superior levels. Information from the feature level then feeds to the letter level, which feeds to the word level; with each node becoming activated whenever there is enough input received to suggest its presence. Concurrently, information can be fed down from the higher levels of processing to activate nodes for words that might occur in that context. When there is enough activation at the word level to activate a specific word node, information then feeds back down to the letter and feature level by either inhibiting or activating specific feature or letter nodes. For example, if a word node for "trip" becomes activated, then there will be some feed back to the letter level with activation of the nodes for the letters "t", "r", "i", and "p". This process can also occur when the word is not correct, thus activating nodes for letters and features that are not there. In addition, sensory information continues to filter through and to add activational input to the nodes of the feature and letter levels. When a feature is activated, it can also send inhibitory input to

Figure 1. McClelland and Rumelhart's (1981) interactive activation model



Note: The arrows in the diagram represent excitatory connections. Circular ends represent inhibitory connections. Intralevel inhibitory loops represent lateral inhibition in which incompatible nodes at the same level inhibit each other.

From "An Interactive Activation Model of Context Effects in Letter Perception: Part 1. An Account of Basic Findings" by J. L. McClelland and D. E. Rumelhart, 1981, Psychological Review, 88(3), 375-407. Copyright 1981 by the American Psychological Association.

the nodes of features that are incongruent with it. This process can reduce the activation of words and letter nodes that had inappropriately been suggested by the context.

Based on Reicher's (1969) results and with their model, McClelland and Rumelhart (1981) suggest that the perception of a letter can be facilitated by presenting it in the context of a word, that our knowledge about words can influence the process of perception. If the same principles apply in medical diagnosis, then a feature presented in the context of a disease should be more easily perceived than if it is presented by itself. Thus, features of a disease should activate nodes for features in a bottom-up manner. If enough of these features nodes are activated, this will activate disease nodes at a higher level. There may be one disease node or many that are activated. There will also be feedback which will activate certain features nodes consistent with the diseases. If the features are not present, there will be a lack of activation coming up from the sensory information, as well as inhibitory input from features incongruent with these absent features. However, because features can be ambiguous, the sensory information activation of the nodes for possibly abnormal features might not be as strong as it would be for unambiguously abnormal features. The activation from ambiguous features could be boosted from the feedback from the disease level. Manipulations, such as priming, should selectively activate certain disease nodes which would then feed back to the feature level. Since there are no other disease nodes activated by the priming, there should be no inhibitory information feeding down from the disease level. The activation of these features nodes should then be very strong. Thus, diagnosticians whose degree of belief in a diagnosis was increased by the manipulation should then also be more likely to call the features consistent with this diagnosis.

If, as suggested by the interactive-activation model, the degree of belief in an hypothesis can affect the detection of features, then another theory which may be of relevance to the field of medical diagnosis is Support Theory. This theory suggests that manipulating the number of hypotheses that a person is asked to evaluate can have a significant effect on the degree of belief in a given hypothesis. Tversky and Koehler (1994) proposed Support Theory as a possible explanation for the finding that people show consistent overconfidence in both the truth of their beliefs as well as in the accuracy of their judgments (Mehle, Gettys, Manning, Baca, & Fisher, 1981). Tversky and Koehler based this theory on the idea that people are overconfident because they neglect to consider other viable alternatives when making judgments.

A key argument underlying their theory is that judged probability estimates are based on the descriptions of events and not on the events themselves. The same event, described in two different ways, could have different probability estimates. Thus, their theory is based on perceived support for a *hypothesis* as described rather than the actual frequency of the occurrences of an *event.* So, instead of measuring the judged probability of an event, A', in terms of its relative frequency; that is, in terms of the standard probability formula:

$$P(A'/B') = \frac{f(A')}{f(A') + f(B')}$$

they argue that researchers should define such a probability in terms of perceived support for one hypothesis as described relative to other hypotheses as described:

$$P(A/B) = \frac{s(A)}{s(A) + s(B)}$$

where s(A) and s(B) refer to perceived support for hypotheses A or B.

A second central assumption of Support Theory is the principle of unpacking: The greater number of alternatives raised in the description of an event, the greater its judged probability will be. A conventional probability function would define the frequency of an event, A', to be equal to its unpack form, (B' v C'):

$$f(A') = f(B' \vee C').$$

However, Tversky and Koehler (1994) argue that when a more detailed description of an implicit hypothesis is presented, its judged probability will increase because it brings to mind previously unthought of alternatives. Thus,

$$s(A) < s(B \vee C),$$

where A is the implicit hypothesis (Ann majors in a natural science), and (B v C) is the explicit disjunction, or unpacked version, of A (Ann majors in either a biological or physical science). To test this assumption, they asked their subjects to evaluate the probability of deaths from a variety of causes. Some of their subjects were asked to evaluate an implicit hypothesis such as "What's the probability that a randomly selected person has died in the previous year due to a natural cause?". The remainder of their subjects were asked to evaluate the same hypothesis, but in its unpacked form ("What's the probability that a randomly selected person has died in the previous year due to heart disease, cancer, or some other natural cause?"). What they observed was that the probability rating for the unpacked version was higher than for the implicit form of the hypothesis (73% vs. 58%, $\underline{p} < .05$).

In a related paper, Koehler (1994) also observed a decreased confidence in judgments made by subjects who generated alternatives as opposed to subjects who evaluated them. He argued that this occurred because the subjects who generated the alternatives probably unpacked more than those who were simply asked to evaluate the same alternatives.

This theory was also applied to the field of medical diagnosis by Redelmeier, Koehler, Liberman, and Tversky (1995). These researchers found that the unpacking effect could also be observed with physicians. Following a brief case description, they asked one group of physicians to rate the probability of each of two diagnoses and a

residual category ("none of the above"). They asked another group to rate the probability of the same two diagnoses, but they unpacked the residual category into four additional categories (3 diagnoses and the "none of the above"). They observed that the average probability
assigned to the residual in the packed group was smaller than the sum of its components in the unpacked group. Again, the argument for the effect of unpacking is that it brings to mind possibilities previously unthought of.

Since the WSE provides evidence that changing the degree of belief in the hypothesis can have an effect on the detection of features, and since Support Theory offers a concrete manner in which we can manipulate the degree of belief in an hypothesis (other than by priming), then it is of interest to investigate whether offering multiple alternatives would make a diagnostician report features differently. If the diagnosticians are told that they are looking at an example of a given disease (the "focal" disease), then they should list many of the features of that diagnosis, since the priming would have strongly activated that one disease node and the activation feedback would serve to activate feature nodes of that disease and to inhibit other feature nodes not consistent with that disease. On the other hand, if several diagnoses are suggested, then multiple disease nodes should be activated, thus changing the nature of the feedback to the feature level. For the features consistent with the focal (or correct) disease, there may be more inhibitory feedback from the additional diseases. Features consistent with the other diseases would receive more top-down activation, leading to them necessitating less activation from the sensory information in order for a feature to be declared present. As for the features of the focal disease, they should receive more top-down inhibition from the additional diseases which are incongruent with them. Thus, these feature nodes would necessitate more activational input from sensory information to reach the criterion point. Also, because generation of hypotheses is believed to bring about unpacking, the performance of diagnosticians who are asked to generate their own hypotheses should resemble that of the unpacked version diagnosticians.

Although it seems likely that the unpacking effect observed by Tversky and Koehler (1994) should also be observed in medical diagnoses and should produce changes at the level of feature detection, studies from Patel and her colleagues (Patel & Groen, 1986; Patel, Groen & Arocha, 1990) suggest that these effects may only be observed in novice diagnosticians. In her widely cited studies, she argues that experts produce a diagnosis using forward reasoning. In other words, diagnoses are driven by bottom-up processing. By this view, data acquisition (detection of features) occurs prior to and seemingly independently of an eventual diagnosis. She based these conclusions on the fact that experts who arrive at a correct diagnosis use, in their verbal reports, propositions in which the direction of reasoning is forward.

There are, however, problems with her method which render it insufficient to prove her argument. The main argument against her conclusions is that, although verbal reports may be an indication of underlying processes, there is no way of knowing when in the process of reasoning these reports are generated. That is, her method does not eliminate the possibility that they are rationalizations of the diagnosticians' conclusions.

In addition to the methodological problems, Patel herself admits qualifications of her conclusions. She cites experiments that can easily be interpreted to imply that experts do not generate the diagnostic hypotheses in an exclusively bottom-up direction. One study she cites is one that Lesgold and his colleagues report in a technical report (in Patel & Groen, 1986), in which they showed that expert behavior is characterized by rapid recognition for cases, something which would not be consistent with a purely forward direction of reasoning.

There is additional evidence to suggest that diagnosticians do not use purely bottom-up processing in making a medical diagnosis. In fact, studies that have been undertaken in medicine suggest that the observation of features is not a stage of clinical investigation which proceeds independently of the diagnoses entertained by the physician or even of the way that the raw evidence is described (interpreted).

Studies in radiology have provided evidence that providing prior information such as the location of tenderness and swelling for the detection of fractures (Berbaum et al., 1988) as well as the tentative diagnosis for the detection of radiographic abnormalities (Berbaum et al., 1986) brings about an increase in the true positive rate of detection of these fractures and lesions. One problem with these studies in radiology, as mentioned by Norman, Brooks, Coblentz and Babcook (1992), is that in all cases, the case history was always consistent with the final diagnosis. Thus, this provided a bias in favor of an abnormal diagnosis for positive films and a bias for a normal diagnosis for negative films. The problem lies in the fact that it remains unclear whether or not this increase results from an increase in the detection of features on abnormal films and of discounting on normal films, or simply from the incorporation of additional information into the overall judgment. In order to address this problem, Norman and his colleagues (1992) ran a similar study, but also did a cross-over of history and radiographs, such that a normal history was not always matched with a normal radiograph and vice-versa. Their subjects were asked to interpret chest radiographs which were either unambiguously normal or abnormal (obvious cases of bronchiolitis), or equivocally normal or abnormal. These radiographs were matched with normal and abnormal case histories. It was found that the clinical histories affected the ratings of features present in the equivocal radiographs.

These above-mentioned studies all suggest that diagnosticians do not proceed in a strictly forward reasoning direction. One problem with the evidence presented above is that one could argue that diagnosticians in the field of radiology may be conditioned to function in a manner which would strongly be influenced by tentative diagnoses. Typically, physicians in this field see patients after much off the clinical data has been accumulated and a tentative diagnosis has been made. Because of the wealth of information that can be gathered prior to them seeing the patients, they may have experienced few cases where the results of their analyses were significantly incongruent with the tentative diagnosis provided by the clinical information. For this reason, the tentative diagnoses presented in the previously mentioned studies may have been fairly persuasive to the diagnosticians. Another problem with the studies mentioned above is that in Norman et al's (1992) study, the disorder selected (bronchiolitis) is one in which radiological findings are particularly suspect. In this case, the strong influence of history on diagnosis may be due to the high level of ambiguity in the films. For these reasons, it would also be of great interest to study the way that diagnosticians reason on the first encounter with a patient, before any information has been gathered and where the presenting signs are fairly straightforward. Are the radiologists behaving in the observed manner because of a predisposition brought about by their field, or because that is the normal way of reasoning of all diagnosticians?

One study that partially answers this question was reported by Norman, Brooks, Shali, Marriott and Regehr (unpublished study). Their study, consisting of both medical students and academic general internists, was aimed at investigating the role of both verbal and visual information in the diagnostic process and at investigating whether or not the way features were described could affect what was detected in patients. Their subjects were given information on 15 cases in 3 passes, with each pass providing them with additional information (either visual or verbal). On pass 1, the subjects were given a short case history. After receiving this history, they were asked to list the diagnostic possibilities in order of likelihood. They did so for each of the 15 cases. Following this, they went through all of the 15 cases again, while receiving additional information. In this second pass, in addition to receiving the case histories, the subjects were shown head and shoulder pictures of each patient. As in the first pass, they were asked to list the diagnostic possibilities. In addition, they were asked to list any clinical features that they saw in the picture. On pass 3, the subjects were again given the case history and the picture. On this pass, they were also given an interpretative description of the clinical features present in the picture. Again, the subjects were asked to give their diagnoses. For half of the experts however, the task on pass 3 was reversed. They were given the diagnosis and then asked to list the features they saw in the pictures. The results of the study show a near linear increase in accuracy across the three passes, for both the students and the experts. An important implication of these results is that features are not self-evident, they do not describe themselves. In this case, the way these features were described was critical to the cueing of the disease.

These authors also observed an interaction between the pass and expertise when comparing diagnostic accuracy from pass 1 to pass 2. This is suggestive that the experts gained more information from looking at the pictures than did the students. More interestingly, these researchers observed that the experts who were given the diagnosis called more of the interpreted features from the pictures than did the experts who where not given the diagnoses (102 vs. 83, $p < .05$; total number of features to be called = 162).

Further qualitative analysis of their data suggests that the experts who were given the diagnosis on pass 3, when compared to the experts who were not given the diagnosis, appeared more likely to use textbook terminology ("periorbital edema" as opposed to "puffy eyes") and to list additional features consistent with the disease but not in the pictures.

The results from this last study suggest that it is not a bias unique to radiology for the diagnosticians to be influenced by tentative diagnoses or by priming. However, the priming done by Norman and his colleagues (unpublished study) was at the level of feature description. While their results indicate that features are not self-evident and that their description can be critical to cueing the diagnosis, they do not permit us to know whether or not the reverse manipulation, cueing of the diagnosis, will also cause variability in feature detection. Although Berbaum et al's (1986) study with radiographic abnormalities does suggest that tentative diagnoses do cause an increase in the detection of lesions, it does not provide us with any information on whether or not this effect is due to a response bias or to an increase in discrimination.

The first experiment presented in this paper is aimed at further investigating this relationship between context and feature detection. There are three hypotheses as to the nature of this relationship.

The first hypothesis is that what occurs during the diagnostic procedure is **independent data acquisition**. Models based on Bayesian logic seem to imply that this is what occurs in diagnosis; that features are self-evident and that diagnostic variability comes from the weighting and organizing of these features. If this is true, then features are detected independently of each other and independently of the diagnostic hypotheses. If this were the case, a diagnosis would be made only if enough features were called to satisfy some decision criterion. Furthermore, information at the disease level would not feed back down to influence perception at the level of features. If this is the type of processing that occurs in the process of making a diagnosis, then cueing a disease should not have any effect on feature calls. Diagnosticians who are manipulated into changing their degree of belief in a hypothesis should not call more features specific to a given disease or even use different terminology than those who are not cued as to the disease. As mentioned, this hypothesis has already been discomfirmed for radiologists (Berbaum et al., 1986; Berbaum et al., 1988; Norman et al., 1992)) and a result opposing this hypothesis appeared in a study with general internists (Norman et al., unpublished study). However, it is important to converge on this latter result with a different manipulation that is also common in medicine: suggested alternative diagnoses.

If such an experiment does in fact disconfirm the hypothesis that feature detection is unaffected by the strength of belief in a diagnosis, there are two ways in which the effect could take place. One possible effect is that presenting diagnosticians with a hypothesis could produce a **pure response bias** by which they call any feature consistent with the diagnosis in mind. That is, the information coming down from the disease nodes would activate all the features nodes consistent with it to a level which reaches a criterion point. It would activate them enough so that additional sensory information activation was not necessary for a feature to be judged present. Thus, once they had a diagnosis in mind they would call all features consistent with the disease, regardless of whether or not these features were present.

The other potential effect of having a diagnosis in mind is that this top-down constraint may **interact with the acquisition of bottom-up information.** While having a diagnosis in mind may render the diagnosticians more likely to call features consistent with this diagnosis, the effect may not be so strong as to prevent discrimination. Thus, if there is an interactive process that is occurring in the manner suggested by McClelland and Rumelhart's interactive activation model, diagnosticians for whom our manipulations have increased the degree of belief in an hypothesis should be more likely to call features that are consistent with the disease. However, the top-down activation would not be so strong as to reach the criterion point without any additional activation from sensory information. Because of the lack of additional activational input from the sensory information, the diagnosticians should not be so biased as to call features which are not present. Thus, manipulations increasing the strength of belief in a hypothesis would increase the likelihood of calling features consistent with it, but the discrimination (from sensory information) would make it so that only features that are present (weakly or strongly) would be called.

The first experiment presented in this paper was aimed at discriminating among these three hypotheses. More precisely, the aim was to disconfirm the hypothesis of independent data acquisition and to investigate whether or not the top-down influence on data acquisition was moderated by discrimination.

Experiment 1: Medical Diagnosis

Experiment 1 was run in order to test the predictions of the three hypotheses mentioned in the introduction. This experiment was run with medical students, since they are more readily available than expert general internists. The first manipulation was to prime a diagnosis by asking the subjects to evaluate a given diagnosis as opposed to asking them to only list the features present (without arriving to any diagnostic conclusion). The second manipulation involved using the principle of unpacking to observe whether or not the number of diagnostic alternatives in mind could affect both strength of belief in a given diagnosis and feature detection.

In order to measure the effects of these manipulations, the rating given to suggested or generated diagnostic hypotheses, the nature of the feature listing, as well as the ratings of the strength of presence of features provided by the experimenter were measured. The measure of disease rating was done as a manipulation check, to ensure that the subjects believed the diagnoses suggested to them and to ensure that the unpacking manipulation worked in the desired direction. The nature of the feature listing was looked at in two ways. First of all, the number of features called that were consistent with the disease was measured to observe if having one as opposed to none, many, or generated diagnoses in mind caused the subjects to call more features of the focal disease. In addition, the terminology used by the subjects was examined to see if we could reproduce Norman et al's unpublished results that subject used more disease specific terminology when they are suggested the diagnosis. Finally, because changes in decisions about the strength of a given feature may not be limited solely to whether or not it came to mind, we looked at the rating the subjects gave to suggested features.

Method

Subjects

Fifty-eight medical students participated in this experiment on a voluntary basis. All of the subjects were in the unit 5 of the McMaster University Medical School (the last pre clinical unit in the undergraduate program). These subjects were chosen because although they are novices in comparison to experts, they should have enough medical training to perform above chance level on the cases presented to them. The students were run in their tutorial groups, up to six at a time.

Materials

The materials for this study consisted of 16 head and shoulder photographs of patients, presented to the students in individual photo albums. Also, each student received a questionnaire to fill out. The questionnaires contained the short case histories as well as the questions regarding each case. The photographs as well as the case histories were the same as those used in Norman et al.'s unpublished study. A brief description of each case can be found in Appendix A.

Design and Procedure

Each tutorial group (consisting of 4-6 students) was assigned, by block randomization, to one of four conditions. The first condition was the "primed" condition, in that they were suggested one diagnostic possibility (1A group). For each of the 16 cases, these subjects looked at a picture of a patient and read a short case history. Following this, they were told that the patient had a condition suggestive of a given disease. They were then asked to list all clinically important features that were present in the picture. Once they had done so, they were asked to rate, on a scale of 1 to 7, the likelihood that the patient had the suggested diagnosis (see Appendix B for sample). On a following page, the subjects were asked to rate, on scale from 1 to 5, the likelihood that the patient had each of 6 suggested features (see Appendix C for sample). The subjects repeated this procedure for each of the 16 cases. For each case, there were three type of features that the students had to rate: the *strong* features, which were consistent with the disease and in the picture; the *weak* features, which were consistent with the disease yet not visible in the picture; and the *control* features which were neither consistent with the disease, nor visible in the picture. These last features were included in order to determine whether the simple act of asking students to rate the strength of a feature would be enough for them to rate it as being present. The nature of each feature presented to the subjects was defined by an expert in internal medicine. In addition to listing features consistent with the diseases, this expert indicated which features were visible in the picture (strong features) and those that weren't (weak features). For each case, the expert also generated a list a features inconsistent with the disease and not in the picture (control features).

For the groups in the 5 alternative condition (5A), the procedure was the same with the exception that they were told that the case history was suggestive of 5 possible diagnoses. The alternatives suggested to them were selected by including the diagnoses which were given most frequently by the students in Norman et al's unpublished study. To encourage further unpacking, the students were also asked to list any additional diagnoses that came to mind. Also, when they were asked to rate the likelihood of the suggested diagnosis, they had to rate the likelihood of each of the five suggested alternatives (see Appendix D for sample).

The groups in the generate condition (GEN) also followed the same procedure of the two previous conditions, with the exception that they were asked to generate the diagnostic alternatives and to rate their likelihood (see Appendix E). Although the questionnaire was set-up so that they had space to list three diagnoses, the students were told that if they could only think of one diagnostic, they should move on (because of time restraints). They were also told that if they could generate more than three alternatives, they should simply write them down and write the number indicative of their degree of belief in that alternative.

As for the tutorial groups placed in the no-diagnosis condition (ND), they were not given the short case histories given to the previous groups. They also received verbal instructions asking them to avoid arriving at any kind of diagnostic conclusion. The questions asking them to list all clinically important features as well as those asking them to rate suggested features were the same as for the three previous conditions (see Appendix F).

Results

Disease Ratings.

The disease ratings given by the subjects in the 1A, 5A and the GEN conditions were analyzed as a manipulation check. The expectations from Support Theory were that the subjects in the 1A groups would give the higher ratings to the focal diagnoses than those in both the 5A and the GEN conditions. The mean ratings given to the focal diagnoses by each group are presented in table 1

A 3 X 16 ANOVA was run on the disease ratings. The between-subject variable was the condition (1A, 5A. GEN) and the within-subject variable was the case (Cases 3, 4, 7, 9, 11, 12, 13, 16, 17, 18, 19, 21, 22, 23, 25, 27; the case numbers are arbitrary and used as the names of the cases). The cases were included in this analysis because they were expected to be heterogeneous. Included them permitted the accounting of the variance due to this heterogeneity. There was no main effect of condition on rating scores, F(2, 24) = .932, p = .407. There was a main effect of cases, F(15, 360) = 3.636, p < .05. There was no significant interaction between the conditions and the cases, F(30, 360) = 1.466, p = .057.

The direction of change in the rating of the focal hypothesis from the 1A to the 5A conditions, as well as from the 1A to the GEN conditions was positive (see table 1).

Table 1

Mean rating scores given for each focal diagnosis as a function of group

	1A	5A	GEN
Case 1	4.9	5.8	5.3
Case 2	4.3	4.1	5.3
Case 3	5.2	5.5	4.9
Case 4	6.5	6.1	6.4
Case 5	5.1	5.6	4.8
Case 6	6.1	6	6
Case 7	5.1	6.4	5.5
Case 8	6.1	6.3	5.4
Case 9	6	6.4	6.5
Case 10	5.4	5.8	5.8
Case 11	5.2	5	5.5
Case 12	5.4	6	5.5
Case 13	4.5	5.5	5.7
Case 14	5.3	5.3	4.3
Case 15	4	4.1	6
Case 16	5.1	5	6
AVG.	5.26	5.56	5.56
ST. DEV.	.68	.72	.58

Note: The scores are given on a 7 point scale. A rating of 1 indicates that the disease is rated "highly unlikely", and a rating of 7 indicates that the disease is rated "highly likely".

When tested against a null hypothesis, the change from the 1A to the 5A condition was significant ($z = 2.45$, $p < .05$). The changes in disease ratings from the 1A to GEN conditions ($z = 1.57$, $p = 0.58$) as well as from the 5A to the GEN conditions ($z = 0.00$, $p = 1.00$) were not significant .

Post-hoc observation of the data indicated that for the cases where the rating of the focal hypothesis went up from the 1A to the 5A condition, there seemed to be a large gap between the rating given to the focal hypothesis and the one given to the next most plausible alternative (in the 5A condition). That is, the focal hypothesis was clearly dominant. To test whether or not the relationship between the degree of focal hypothesis dominance and the direction of change was significant, a Pearson product-moment correlation test was run on the two variables. While the correlation between the two was positive, the relation between the two variables was not significant $r(16) = .099$, $p = .715$. (see figure 2a for scatter plot). Because the top four cases seemed to be behaving differently than the other 12 (possible ceiling effect and strong effect of regression to the mean) they were removed and the Pearson product-moment correlation approached significance $r(12) = .559$, $p = .059$. (see figure 2b for scatter plot)

Feature ratings

The ratings of the strength of presence of the suggested features was analyzed for the four conditions. Again, it was expected that the primed condition (1A) would rate features consistent with the focal disease as more likely to be present than the subjects in the unprimed condition (ND). In addition, it was expected that the 5A and the GEN conditions would rate these features as more likely than subjects in the ND condition, but less likely than the subjects in the 1A condition. Depending on which of the initial hypotheses is true, we expected a different pattern of results. If the effect of

Figure 2. Scatter plot of the change in focal disease rating as a function of the degree of focal hypothesis dominance for 16 cases (2a) and with the top 4 cases removed (2b): Medical cases.

Figure 1a.

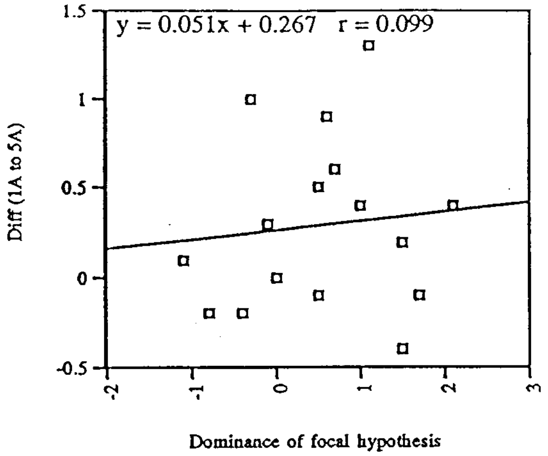
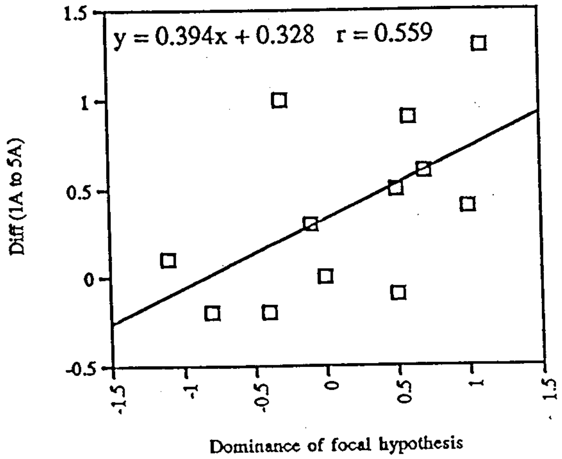


Figure 1b.



Note: The difference in the rating of the focal hypothesis from the 1A to the 5A condition [Diff (1A to 5A) is plotted as a funtion of the difference in rating between the focal and the next most plausible alternative [Dominance of focal hy]. A positive value for "Diff (1A to 5A)" indicates that the focal hypothesis was rated more likely by the 5A group than by the 1A group. A positive value for "Dominance of focal hy" indicates that the focal hypothesis was rated as being more likely than the next most plausible alternative.

having a diagnosis in mind causes diagnosticians to call anything consistent with the disease, then both the weak and the strong features should be rated as likely by these diagnosticians. If however, the influence of having a diagnosis in mind is moderated by discrimination, then these diagnosticians should only give higher ratings to the strong features.

The mean ratings of strength of presence of the features given in each condition are presented in table 2. A 4 X 3 ANOVA was run on the ratings of the strength of presence of the suggested features. The between-subject variable was the condition (1A, 5A, ND, GEN) and the within-subjects variable was the type of features (strong, weak, control). There was no significant effect of condition on feature ratings, $\underline{F}(3, 115) = .289$, $p = .833$. There was a significant effect of type of feature, $\underline{F}(2, 230) = 235.04$, $p < .05$. There was no significant interaction between the condition and the type of feature, $\underline{F}(6, 230) = .441$, $p = .851$. A post-hoc analysis (Tukey HSD) indicated that the strong features were rated a being significantly more likely than the weak (4.25 vs. 2.99, $\underline{p} < 0.05$) and the control features (4.25 vs. 2.44, $\underline{p} < 0.05$). The weak features were rated as being more likely than the control features (2.99 vs. 2.44, $\underline{p} < 0.05$). This analysis acted as a confirmation of the expert's judgment of the presence or absence of the features in the pictures.

Because the results of disease ratings did not come out as expected, we decided to see if there was a relationship between the degree of belief in the focal hypothesis and the rating given to features of a disease. Regardless of the manipulations, subjects who have a higher degree of belief in a hypothesis would be expected to have a higher degree of belief in the presence of features consistent with it. However, a Pearson product-moment correlation indicated that there was no significant relationship between the rating of the likelihood of the focal hypothesis and the rating of the presence of the

Table 2

Mean ratings of strength of presence of suggested features as a function of group

	1A		5A		ND		GEN	
	AVG	SD	AVG	SD	AVG	SD	AVG	SD
Strong Features	4.37	.09	4.26	.10	4.24	.09	4.12	.11
Weak Features	3.08	.13	2.96	.13	2.95	.13	3.01	.12
Control Features	2.33	.13	2.48	.12	2.47	.13	2.40	.11

AVG= average SD= Standard Deviation

Note: The scores are given on a 5 point scale. A rating of 1 indicates that the feature is rated "highly unlikely" to be present, and a rating of 5 indicates that the feature is rated "highly likely" to be present.

feature, for either the strong features, r(97) = .123, p = .231; the weak features, r(90) = .124, p = .243; or the control features, r(102) = -.070, p = .48.

Because the degree of focal hypothesis dominance shows a trend of being positively correlated with the direction and size of change in disease rating from the 1A to the 5A condition, we decided to look if this relationship also existed with the ratings of presence of the suggested features. A Pearson product-moment correlation showed there was no significant relationship between the degree of focal hypothesis dominance and the change in rating of presence of the features, for either the strong features, r(32)= .222, p = .067; or the weak features, r(30) = -.008, p = .964. There was a significant positive relationship between the degree of focal hypothesis dominance and the change in strength of feature rating for the control features, r(34) = .446, p = .008.

Feature listing

It was predicted that the subjects in the 1A condition would call more features specific to the focal disease than the subjects in the three other groups, and that they would use more disease specific terminology than the ND group. Analyses were run in order to see if the groups differed in the total number of features called and in the number of disease specific features called. Because there were no observable differences in terms of the terminology used (all the groups used the same words for the features), it was not analyzed. A summary of the total number of features and the proportion of strong and weak features called per case is presented in table 3.

A 4 x 16 ANOVA was run on the total number of features called per case. The between-subject variable was the condition (1A, 5A, ND, GEN) and the within-subject variable was the case (3, 4, 7, 9, 11, 12, 13, 16, 17, 18, 19, 21, 22, 23, 25, 27).

Table 3

Mean number of features and proportion of strong and weak features called per case by each group.

	1A	5A	ND		GEN	
				COR		INC
Mean number of features called	4.17	4.43	3.68		3.64	
Prop. weak features called	.11	.11	.09		.10	
Prop strong features called	.59	.50	.44	.60	.50	.40

There was no significant main effect of condition, F(3, 54) = 1.44, p = .24. There was a significant main effect due to cases, F(15, 810) = 16.07, p < .05. There was no significant interaction between the condition and the cases, F(45, 810) = .73, p = .91.

The proportion of features consistent with the focal diagnoses that were called was also analyzed. One prediction regarding feature calls was that diagnosticians who had the diagnosis in mind should call more features consistent with it than those who do not have the diagnosis in mind. Another prediction was that if the influence of having a disease in mind works without discrimination, then the diagnosticians with the diagnosis in mind should call any features of the disease, regardless of whether or not they are present (call both the weak and the strong features). If this influence is moderated by discrimination, then only the strong features should be called more often by the diagnosticians with the disease in mind.

A 4 x16 ANOVA was run on the proportion of the strong features called per case. The between-subject variable was the condition (1A, 5A, ND, GEN) and the within-subject variable was the case (same as above). There was a significant main effect of condition, F(3, 54) = 6.28, p < .05; and of cases, F(15, 810) = 26.59, p < .05. There was a significant interaction between the condition and the cases, F(45, 810) = 1.75, p < .05. A post-hoc analysis (Tukey HSD) indicated that the 1A group called a higher proportion of strong features per case than did the GEN group (.59 vs. .50, p = .001) and the ND group (.59 vs. .44, p < .05). Although the 1A group showed a trend of calling a higher proportion of strong features than the 5A group, the difference was not significant (.59 vs. .50, p = .13).

There were two types of trials in the GEN condition [those in which the correct diagnosis was generated (GEN COR) and those in which it wasn't (GEN INC)]. Therefore, a secondary analysis (ANCOVA) was done on the GEN condition, with Correct and Incorrect diagnosis used as the dummy variable. There was a marginally significant difference between the two types of trials, F(1, 13) = 3.58, p = .08. As can be seen in table 3, the GEN COR group performed equally to the 1A condition, and the GEN INC group called a smaller proportion of the strong features than the ND group.

To investigate whether the same group differences could be observed with the weak features, a 4 x16 ANOVA was run on the proportion of the weak features called per case. The between-subject variable was the condition (1A, 5A, ND, GEN) and the within-subject variable was the case (same as above). There was no main effect of condition, F(3, 54) = 1.06, p = .37, but there was a main effect of cases, F(14, 756) = 96.31, p < .05. There was no significant interaction between the condition and the cases, F(42, 756) = .58, p = .98.

Discussion

A surprising finding from this experiment was that the rating of the likelihood of the focal hypothesis went up in the 5A condition as compared to the 1A condition. This is in the opposite direction as that predicted by Tversky and Koehler's (1995) Support Theory. In this case, unpacking the implicit disjunction did not cause a decrease in the rating given to the focal hypothesis. One reason for the differing results may be due to a difference in the amount of information provided to the students. In this experiment, although the case histories may have been vague enough to make many diagnoses plausible, it is possible that the students could draw enough information from the pictures to eliminate some of the alternatives, leading to the focal hypothesis being dominant. This is supported by the result of a positive relationship between the degree of focal hypothesis dominance and the change in focal hypothesis rating from the 1A to the 5A condition. Although this relationship is not significant, there are reasons to believe that with a better manipulation of the focal hypothesis dominance, it would be possible to observe a significant relationship between the two variables. The first reason is that when the cases with an initial focal hypothesis rating above 6 were removed (because of a ceiling effect and an extreme influence of regression to the mean), the correlation between the two factors approached significance. Also, because the plausibility of the alternatives was not controlled, there were few cases where the focal hypothesis was extremely dominant. If the plausibility of the alternatives presented with the focal hypothesis was manipulated such that there was a large gap between the rating given to the focal hypothesis and that given to the next most plausible alternative, it could be expected that the rating for the focal hypothesis should increase from the initial rating. Experiment 2 is aimed at investigating whether or not there is a strong positive relationship between the degree of focal hypothesis dominance and the change in focal hypothesis rating from the packed to the unpacked condition.

As for the ratings of strength of presence of the features, the results indicate that these ratings are not correlated with the degree in belief in the disease. These results seem to lend support to the hypothesis that feature detection is independent of the diagnosis that a diagnostician has in mind. However, the results from the analysis on the feature calls do not support this hypothesis. These latter results actually suggest that having a diagnosis in mind causes subjects to call more of the features of the disease than not having it in mind. The differences between the groups are actually quite impressive. At one extreme, the subjects who generated their own hypothesis and those to whom it was suggested called 60% of the disease consistent features. At the other extreme, the subjects who did not generate the correct diagnosis only called 40% of these features, while those who were asked to avoid arriving at any diagnostic conclusion called 44% of these features. Therefore, our results show that having one diagnosis in mind can cause novice diagnosticians to call up to 20% more disease-specific features than not having it in mind. The results also indicate that this effect is moderated by the principle of unpacking, as the subjects in the 5A group called a proportion of features (50%) which was between the 1A and GEN COR groups, and the ND and GEN INC groups. This effect is not due to any group simply calling more features than another as there were no differences in terms of the total number of features called.

In addition to supporting the hypothesis that feature detection is influenced by context, these results also indicate that asking subjects to rate the strength of presence of features is insufficient to observe this influence. A potential explanation for this may be that all of the subjects see the same elements in the pictures, but that seeing them in the context of a disease may push an element from a state of natural variation to one of an abnormal feature. While most subjects may notice that a patient has a pale complexion, those that have the diagnosis of "leukemia" in mind may actually "see" the pale complexion as pallor due to anemia. While it might be expected that if this were occurring, those who did not have the diagnosis in mind would mention the pale complexion, that may not happen. Because the subjects were asked to list all the **clinically** important features, they may have judged a pale complexion to be **clinically** unimportant, thus not have called it. However, if this were occurring, the subjects who do not call the features should also rate their strength of presence lower than those subjects who did call the features. This was not observed in the results of feature ratings.

A more likely explanation for these results may be that the effect of having a disease in mind on feature detection is one of focus of attention. Having a diagnosis in mind may act to guide the attention of the subjects to specific features. Subjects who do not have the hypothesis in mind would be doing a more scattered search of features, while subjects who have many diagnoses in mind would have their attention distributed by looking for features of a number of diseases. Suggesting features to the subjects may have the same effect as having the diagnosis in mind by focusing their attention on these features. Differences between the groups would then be erased.

In addition to investigating whether or not feature detection was influenced by the diagnostic context, we wanted to investigate whether or not this influence consisted of a response bias or of an increase in discrimination.

If the manipulation had an effect of response bias, then it was predicted that the subjects would call the features of the diseases regardless of whether or not these features were present in the pictures. To investigate whether or not they were doing so, we analyzed the proportion of weak features called by the subjects. As mentioned, the weak features were the features that were consistent with the disease, but not present in the pictures. For all conditions, these features were rated as less likely to be present than the strong features. In fact, the presence of these features was rated, on average, as being "uncertain", while the presence of the strong features was rated as being, on average, "likely" to "very likely". In addition, the subjects self-generated a smaller proportion of the weak features than the strong ones. Only 9-11% of the weak features were self-generated by the groups. These numbers come from the fact that there were 2-3 features which were generated by many of the subjects. In addition, while the diagnosticians who had a diagnosis in mind (1A) called a higher proportion of strong features than those who didn't (ND, GEN INC), this difference was not observed with the weak features.

One result which was not observed in this study was the change in terminology used by the diagnosticians who were cued the diagnosis. While Norman et al (unpublished study) found that physicians to whom the diagnosis had been cued used more disease-specific terminology that those diagnosticians who did not have the diagnosis in mind, this result was not observed with medical students. The students all used the same terminology when they self-generated the features. The reason cannot be ignorance of the technical terms because in every case where the disease relevant features were self-generated, the terminology used was disease-specific. There are two potential explanations for the lack of changes in terminology. One reason may be that the students are "closer" to the textbooks than the experts. Because they are still learning the formal relationships between diseases and features, they are most probably relying heavily on textbooks. Their context of learning is probably heavily filled with terminology from these textbooks, thus making them prone to using this specialized vocabulary. Another potential explanation is that we may have biased the use of terminology of the students with our experimenter-suggested features. In the lists of features that the students had to rate for strength of presence, the terminology was always disease-specific ("periorbital edema" as opposed to "puffy eyes"). This may, in some way, have pushed the students towards using textbook terminology. As it stands, there is no way to know for certain which of the potential explanations is the correct one.

Experiment 2: General knowledge

This experiment was run in order to investigate whether or not the increase in rating of the focal hypothesis from the packed condition to the unpacked condition is related to the degree of focal hypothesis dominance. Remember that based on the predictions of the support theory, subjects in the 1A condition should have rated the focal hypotheses as being more likely than would have the subjects in the 5A group. That is, the unpacking should have brought about a decrease in the belief in the focal hypothesis. We found the opposite results. We also observed that the degree of focal hypothesis dominance was somewhat related to the direction of the change in disease rating. This second experiment was run in order to test this relationship more rigorously. If the increase of the rating of the focal hypothesis from the 1A condition to the 5A condition is due to a large dominance of the focal hypothesis, we should be able to obtain the same results by manipulating the plausibility of the alternatives. That is, if the focal hypothesis is presented with other alternatives that are plausible, we would expect to observe the unpacking effect. If however, the focal alternative is presented with some unlikely alternatives, then the focal hypothesis should be more clearly dominant than with plausible alternatives. In this case, it would be expected that the degree of belief in the focal hypothesis would increase from the one alternative condition to the multiple alternative condition. As undergraduate psychology students were more readily available than medical students, this experiment was run using them. Because medical diagnosis questions would be beyond their level of instruction, we decided to use general knowledge questions.

Method

Subjects

Fifty-two undergraduate psychology students participated in this study in return for course credit. Because of the verbal nature of the stimuli, it was required that their first language be English.

Materials

The questions used in this experiment were adapted from Nelson and Narens' (1980) general knowledge questions. Their questions involved the recall of the correct answer. Because our procedure was looking at recognition, the questions chosen were the ones that received the lowest accuracy scores and for which alternatives of varying degrees of plausibility could be generated.

Design and procedure

The variables manipulated in this experiment were the number of alternatives presented as well as the plausibility of the alternatives presented with the focal hypothesis. The three levels of plausibility were the plausible alternatives, the unlikely alternatives, and the not in domain alternatives (alternatives that are unrelated to both the question and the focal hypothesis) (see Appendix G for sample). For each question, the alternatives were always presented with the option "something else". There were 52 questions, each presented either in the 1A condition (focal hypothesis + something else), or in the three alternative (3A) condition (focal + either plausible, unlikely or not in domain alternatives + something else). The alternatives presented with the focal hypothesis were always of the same category of plausibility. That is, a plausible alternative was never presented with an unlikely or not in domain alternative, and vice-versa. The 4 versions of the questions were counterbalanced so that a subject did not answer two forms of the same question. In each of the 4 counterbalancing order, the subjects first gave their degree of belief for 13 questions in the 1A form. They then gave their degree of belief for the 39 remaining questions in either of the 3A forms (randomly assorted). These will be called the between-subject questions.

Also, we wanted to see if the degree of belief in the focal alternative could be changed within a session. Therefore, the initial 13 questions were repeated at the end of the questionnaire in their 3A form. The type of alternatives presented with the focal alternative was randomly selected, with the restriction that there had to be approximately the same number of plausible, unlikely and not in domain questions. These will be called the within-subject questions.

The subjects were given these questions in the form of a questionnaire that they answered on a individual basis. Subjects were either run individually or in groups of up to 10. They received additional verbal instructions telling them that the suggested answers were mutually exclusive (thus should sum to 100%) and that the correct answer may or may not be in the suggested alternatives (to avoid having them rate the most likely alternative as 100% likely)

Results

Because the students' judgments of plausible and unlikely alternatives was not as we had predicted, the questions were analyzed regardless of their initial classification (although an initial manipulation check indicated that the initial groupings acted in the desired direction, see table 4).

A regression analysis was done on both sets of data to determine if there was a positive relationship between the degree of focal hypothesis dominance and the change in the disease rating from the 1A to the 3A condition (independently of their original grouping). The correlation between the two variables was significantly positive for the between-subjects questions, $\underline{r}(166) = .537$, $\underline{p} < 0.000$), and for the within-subject questions, $\underline{r}(58) = .393$, $\underline{p} = .002$ (see figures 3 and 4).

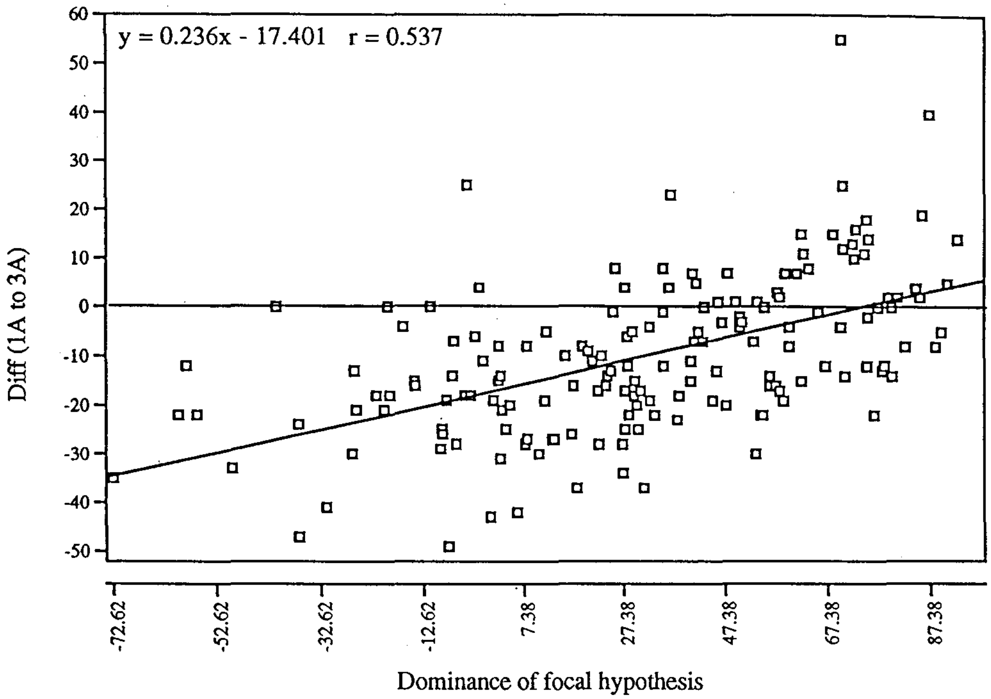
Table 4

Manipulation check for the general-knowledge study, for the between-subject and the within-subjects questions

	Rating of focal hy in 1A condition	Diff in rating between the focal and the next most plausible alternative	Diff in rating of focal hy. from 1A to 3A condition
Between-subject questions			
Plausible	56	16	-17
Unlikely	56	33	-8
Not in Domain	56	45	-5
Within-subject questions			
Plausible	63.4	36.4	-6.48
Unlikely	51	38.5	-1.94
Not in Domain	56.81	55.3	4.87

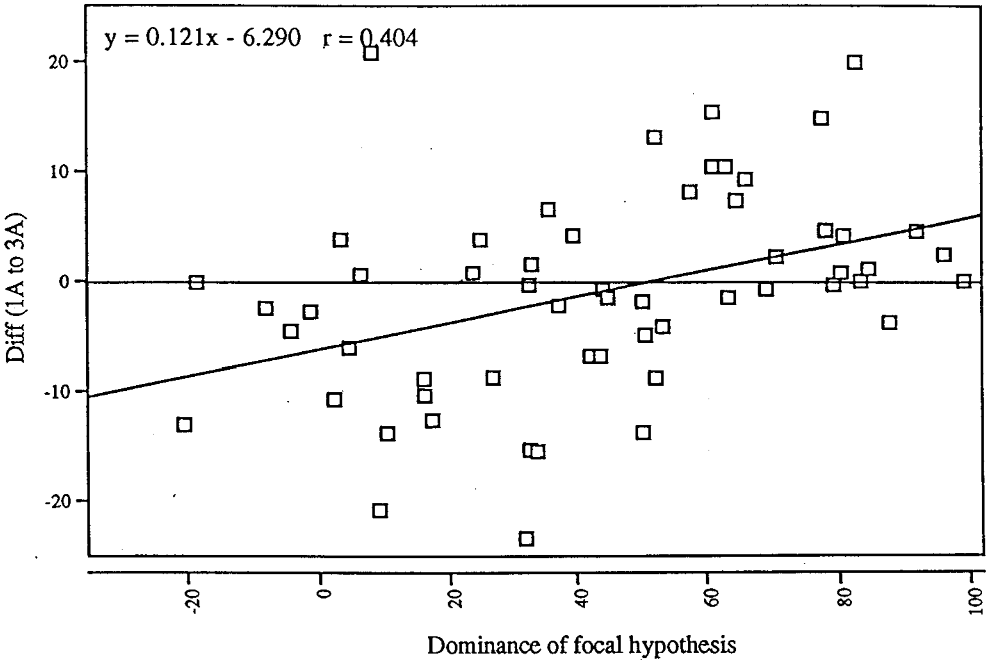
Note: The scores are given in terms of percentage points. For example, for the "plausible" between-subject questions, the focal hypothesis was rated, on average, 16 percentage points higher than the next most plausible alternative.

Figure 3. Scatter plot of the change in focal hypothesis rating as a function of the degree of focal hypothesis dominance for the between-subjects questions: General knowledge study.



Note: The difference in the rating of the focal hypothesis from the 1A to the 3A condition [Diff (1A to 3A)] is plotted as a function of the difference in rating between the focal and the next most plausible alternative [Dominance of focal hy.]. A positive value for "Diff (1A to 3A)" indicates that the focal hypothesis was rated more likely by the 3A group than by the 1A group. A positive value for "Dominance of focal hy." indicates that the focal hypothesis was rated more likely than the next most plausible alternative

Figure 4. Scatter plot of the change in focal hypothesis rating as a function of the degree of focal hypothesis dominance for the within-subjects questions: General knowledge study.



Note: The difference in the rating of the focal hypothesis from the 1A to the 3A condition [Diff (1A to 3A)] is plotted as a function of the difference in rating between the focal and the next most plausible alternative (Dominance of focal hypothesis). A positive value for "Diff (1A to 3A)" indicates that the focal hypothesis was rated more likely by the 3A group than by the 1A group. A positive value for "Dominance of focal hy" indicates that the focal hypothesis was rated more likely than the next most plausible alternative

In addition to predicting a positive correlation between these two variables, the hypothesis of this experiment also predicted that when the degree of focal hypothesis dominance was high, the change in the rating of the focal hypothesis from the 1A condition to the 3A condition should be positive. This prediction was tested with a regression analysis. The difference in the rating given to the focal hypothesis from the 1A condition to the 3A condition (Y) was the dependent variable and the degree of focal hypothesis dominance (X) was the independent variable. The analysis done on the regression line of these two variables tested whether or not the predicted value of Y at the maximum observed value of X was significantly above zero. This analysis was done using a T-test statistic described by Kleinbaum, Kupper and Muller (1988; p.60). The results of this analysis indicate that at the maximum observed value of the degree of focal hypothesis dominance, the difference in the rating for the focal hypothesis from the 1A condition to the 3A condition was significantly above zero for the between-subject questions, $\underline{t}(165) = 2.16$, $\underline{p} < .05$, and for the within-subject questions, $\underline{t}(56) = 2.75$, $\underline{p} < .05$.

Discussion

The results of Experiment 2 confirm the hypothesis that the larger the degree of focal hypothesis dominance, the more likely the change in confidence rating of the focal hypothesis from the 1A to the 3A condition is to be positive. Not only does the degree of focal hypothesis dominance moderate the unpacking effect, but when the degree of dominance is high, it can cause a reversal of the unpacking effect. These results appear to represent a serious limitation to the support theory. That is, in order to have a decrease in the confidence rating of a focal hypothesis with unpacking, the alternatives that are unpacked must be plausible. In other words, the degree of focal hypothesis dominance must be kept at a minimum in order to get the unpacking effect suggested by Support Theory. This would explain the divergence between our results and those of

Redelmeier et al. (1995). While these authors also gave short case histories to their subjects, these histories were vague enough to make several diagnoses equally plausible. By not providing their subjects with any additional source of information (such as the pictures used in our medical diagnosis study), they prevented any possibility of eliminating alternatives leading to dominance of the focal hypothesis. Because the focal hypothesis never became dominant for their subjects, its rating always decreased when presented with other equally plausible alternatives.

General Discussion

The main objective of Experiment 1 was to investigate whether or not feature detection was affected by manipulations which changed the degree of belief in a hypothesis. If these manipulations did affect feature detection, was it by means of a response bias or an increase in discriminability?

The possible relation between the degree of belief in the diagnostic hypothesis and feature extraction was conceptualized as being consistent with McClelland and Rumelhart's interactive activation model (1981). That is, when disease nodes are activated, they feed back activational input to feature nodes consistent with them. However, this process is not purely top-down as the feature nodes need additional activation from the extraction of sensory information in order to reach a criterion point where the features are declared present. The results of the analysis of the self-generated features support the hypothesis. Thus, while having the focal diagnosis in mind did cause the subjects to self-generate more of the disease's features, the influence was not so strong as to make them call all of the disease's features, regardless of their presence. In addition, the unpacking manipulation seemed to cause a dilution of this effect. The subjects who had many diagnoses in mind showed a trend of calling a greater proportion of the disease relevant features than those who did not have the diagnosis in mind, but less than those who only had one diagnosis in mind.

This pattern of increased discriminability was not observed with the ratings of strength of presence of the suggested features. As mentioned in the discussion section of Experiment 1, a potential explanation for these apparently diverging results may be that suggesting features to subjects causes them to focus their attention to these features. While those subjects who already have the focal diagnosis in mind may already have their attention focused on these features, those that have either many or no diagnoses in mind may need this suggestion of features to focus their attention to them. But, once this is done, there is no further effect of having the focal diagnosis in mind.

These results are of importance to medical diagnosis because they indicate that feature detection is not independent of the diagnosis that a physician has in mind. While approaches and computer programs based on Bayesian logic deal with the issue of diagnostic variability due to availability of diagnoses or to ignorance of link between features and disease, they fail to address this issue of variability in feature detection. In doing so, physicians may be misguided in their belief that these computer models are an important help in diagnosis. If a physicians fails to detect features that are present due either to having the wrong diagnosis in mind or of having no diagnosis in mind (thus not knowing what to look for), the correct diagnosis may not be suggested as a possibility by the computer program. It may be of interest to develop systems which, prior to suggesting diagnoses, would ask the physicians to rate the strength of presence of additional features; features that are associated with any feature listed. The ratings given to these features, showed to be independent of the diagnostic context, may change the diagnoses generated by the programs.

The second important result of these experiments is that a limitation to Support Theory (Tversky & Koehler, 1994) has been discovered. The results of Experiment 1 showed that unpacking the diagnostic alternatives actually served to increase the confidence rating given for the focal hypothesis. Based on these results, it was believed that this increase in rating may have been related to the degree of focal hypothesis dominance, that is to the size of the difference between the rating given to the focal hypothesis and the rating given to the next most plausible alternative. The results from Experiment 2 confirm this. Thus, although the unpacking effect is often observed with the explicit disjunction of a residual category, the degree to which the focal hypothesis dominates over the other alternatives may represent an important limitation to Support Theory.

In addition to confirming the relationship between the degree of focal hypothesis dominance and the direction of change in rating, Experiment 2 presents a greater challenge to Support Theory than Experiment 1 because it more closely resembles Tversky and Koehler's (1994) as well as Redelmeier et al's (1995) procedure. The design of Experiment 1 was different from their designs in several important ways: The residual category was implicit, and the students were not instructed that the cases were mutually exclusive, resulting in the total rating for all possibilities often adding to more than 100%. Finally, instead of being asked to give probability ratings, the students were asked to rate the likelihood of each diagnosis on a scale of 1 to 7. These design differences were allowed because there was little initial doubt that the unpacking effect would not occur. Because of this belief, and because of the desire to maintain ecological validity, these changes were believed to be acceptable. However, after Experiment 1, it was feared that these differences may have had some role to play in the difference between our results and those of Tversky and Koehler (1994) and of Redelmeier et al., (1995). For this reason, the design of Experiment 2 more strictly resembled that of these researchers: the residual category was always explicit; the subjects were asked to give their confidence ratings in terms of percentages; and the students were told that their confidence ratings had to sum to 100% for each question (because the possibilities were exclusive). This confirmation of the results of Experiment 1 gives stronger support to the notion that the divergence between our results and those of Tversky and his colleagues is due to the degree of focal hypothesis dominance and not to differences in experimental designs

Future Research

There are a number of studies that can be undertaken as a follow up to these two experiments. As mentioned in the general discussion, the subjects with the focal diagnosis in mind did self-generate more of the features consistent with the disease. There are however, certain issues which are not addressed with our manipulations. In all the cases where the focal hypothesis was suggested to the subjects, the diagnosis was always the correct one. It is not possible to know from these results whether having **any** disease in mind, correct or incorrect, will cause subjects to call more of the features consistent with it. That is, if the subjects are suggested a diagnostic hypothesis that is incorrect, will they call more feature consistent with it, or does the diagnostic suggestion have to be correct?

This leads to another issue that cannot be addressed with these results. It is unclear whether subjects in Experiment 1 self-generated more of the features consistent with the focal diagnosis because they had it in mind or because they believed it to be true. In all cases, the diagnoses suggested to the students were rated as likely to some degree. If the subjects were suggested a diagnosis which they believed to be unlikely, would they also self-generate more features consistent with it? In order to address this question, a study needs to be run in which students are suggested diagnostic hypotheses which they judge to be unlikely. These suggestions could be judged unlikely because they are incorrect or perhaps because the presenting subject is an atypical case of a given disease. If, in this case, the subjects self-generate more features of the disease than subjects who do not have this disease in mind, then it would be possible to state that the influence of suggesting a diagnosis on feature detection is due to having the diagnosis in mind, and not simply of believing this diagnosis to be likely.

Another line of research that could be extended is that regarding Support Theory. While the results of Experiment 2 (general knowledge study) show that a high degree of focal hypothesis dominance does cause an increased belief in it following an unpacking manipulation, it would be of interest to investigate whether this result could be replicated with medical cases. In the general knowledge study, each of the questions had only one clearly correct answer (capital of Australia = Canberra). In medicine, this is not always the case. Because of an increased level of ambiguity in medical diagnosis, it may be difficult to observe high levels of focal hypothesis dominance. Perhaps the only way to observe this is to present focal hypotheses with alternatives such as the "not-in-domain" alternatives of the general knowledge study; that is, to present the focal diagnosis with alternatives that are completely unrelated to the case at hand. If this last manipulation were to increase the degree of belief in the focal diagnosis, this could be an interesting finding, especially if this increase in belief was accompanied by a greater number of features detected. While it appears counter-intuitive for a diagnostician to judge heart failure to be less likely and to call less of its features if it is considered alone than if it is considered with mumps, tonsillitis, and hypothyroidism as the alternatives, the results from the general knowledge study suggest that this is what should occur.

A third line of study which should be considered is to run all of the above studies with expert physicians. It would be of great interest to investigate whether these manipulations would affect experts. Can experts' degree of belief in a diagnosis be changed by priming or by unpacking the alternatives? In addition, if the degree of belief in the diagnostic hypothesis can be manipulated in experts, it would be of interest to investigate whether this would also affect their feature calls. Results from these last experiments would inform us if the influence of diagnostic context on feature detection is a characteristic of novice diagnosticians (perhaps due to limited knowledge) or a basic heuristic of any diagnostic task.

In conclusion, both Experiment 1 of this paper and Norman et al's unpublished study give trouble to the Bayesian approach to medical diagnosis. The results of both studies show that the detection of features is not only affected by the way they are described, but also by whether or not the diagnostician is considering the diagnosis consistent with these features. While the focus of both experiments was the interaction of feature extraction and the diagnostic hypotheses, the manipulation used in each study was different. Norman and his colleagues manipulated the description of the features, while we manipulated the degree of belief in a diagnostic hypothesis. The results of both manipulations converge to suggest that features are not self-evident, they do not describe themselves. Researchers trying to design computer programs to increase diagnostic accuracy should perhaps concern themselves with these issues of feature detection, rather than simply with those of interpretation and combination of these features.

References

Berbaum, K. S., Franken, E. A., Dorfman, D. D., Barloon, T., Ell, S. R., Lu, C. H., Smith, W., & Abu-Yousef, M. M. (1986). Tentative diagnoses facilitate the detection of diverse lesions in chest radiographs. Investigative Radiology, 21, 532-539.

Berbaum, K. S., El-Khoury, G. Y., Franken, E. A., Kathol, M., Montgomery, W. J., & Hesson, W. (1988). Impact of clinical history on fracture detection with radiography. Radiology, 168(2), 507-511.

Berner, E. S., Webster, G. D., Shugerman, A. A., Jackson, J. R., Algina, J., Baker, A. L., Ball, E. V., Cobbs, G. G., Dennis, V. W., Frenfel, E. P., Hudson, L. D., Mangall, E. L., Rackley, C. E., & Taunton, O. D. (1994). Performance of four computer-based diagnostic systems. The New England Journal of Medicine, 330, 1792-1796.

Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. Cognitive Psychology, 24, 411-435.

Kleinbaum, D. G., Kupper, L. L., & Muller, K. E. (1988). Applied Regression Analysis and Other Multivariable Methods, PWS-KENT, Boston: MA.

Koehler, D. J. (1994). Hypothesis generation and confidence in judgment. Journal of Experimental Psychology: Learning, Memory, and Cognition, 20, 461-469.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. Psychological Review, 88(5), 375-407.

Mehle, T., Gettys, C. F., Manning, C., Baca, S., & Fisher, S. (1981). The availability explanation of excessive plausibility assessments. Acta Psychologica, 49, 127-140.

Nelson, T. O., & Narens, L. (1980). Norms of 300 general-information questions: Accuracy of recall, latency of recall, and feeling-of-knowing ratings. Journal of Verbal Learning and Verbal Behavior, 19, 338-368.

Norman, G. R., Brooks, L. R., Coblentz, C. L., & Babcook, C. J. (1992). The correlation of feature identification and category judgments in diagnostic radiology. Memory and Cognition, 20(4), 344-355.

Norman, G. R., Brooks, L. R., Shali, V., Marriott, M., & Regehr, G. (1996). Expert novice differences in the use of history and visual information from patients. Unpublished Manuscript.

Patel, V. L., & Groen, G. J. (1986). Knowledge based solution strategies in medical reasoning. Cognitive Science, 10, 91-116.

Patel, V. L., Groen, G. J., & Arocha, J. F. (1990). Medical expertise as a function of task difficulty. Memory & Cognition, 18(4), 394-406.

Redelmeier, D. A., Koehler, D. J., Liberman, V., & Tversky, A. (1995). Probability judgment in medicine: Discounting unspecified possibilities. Medical Decision Making, 15(3), 227-230.

Reicher, G. M. (1969). Perceptual recognition as a function of meaningfulness of stimulus material. Journal of Experimental Psychology, 81, 274-280.

Sackett, D. L., Haynes, R. B., & Tugwell, P. (1985). Clinical epidemiology: A basic science for clinical medicine. Boston: Little, Brown and Company.

Tversky, A., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. Psychological Review, 101(4), 547-567.

Whittlesea, B. W. A., & Brooks, L. R. (1988). Critical influence of particular experiences in the perception of letters, words, and phrases. Memory & Cognition, 16(5), 387-399.

Appendix A : Summary of each case

3V: Stomach Cancer

Case History: A man came to the emergency room with haematemesis and mild epigastric pain.

Alternative Diagnoses: liver cancer, gastric ulcers, esophageal varices, lung cancer

Strong Features: weight loss, supraclavicular nodes

Weak Features: jaundice, enlarged scalene nodes

Control Features: epicanthal folds, myxedema

4V: Myasthenia Gravis

Case History: A man presented with double vision and dysphagia

Alternative Diagnoses: transient ischemic attack, cerebral-vascular accident, tumor, multiple sclerosis

Strong Features: bilateral ptosis, facial weakness producing a snarling expression when attempting to smile

Weak Features: malar rash, weight loss

Control Features: periorbital edema, round face

7V: Polymyositis

Case History: A lady presented with gradual onset of malaise and weakness associated with an 8 kilogram weight loss

Alternative Diagnoses: cancer, lupus, hypothyroidism, hyperthyroidism

Strong Features: fever, proximal muscle weakness, heliotrope rash

Weak Features: periorbital edema,

Control Features: enlarged scalene nodes, trunkal obesity

9V: Turner's Syndrome

Case History: A woman presented with primary amenorrhea

Alternative Diagnoses: hypogonadism, ovarian failure, Cushing's disease, absence of ovaries or uterus

Strong Features: webbed neck, abnormal facies (narrow maxilla, micromandible, epicanthal folds)

Weak Features: mental retardation, pigmented nevi

Control Features: lymphadenopathy, petichiae

Appendix A (page 2)

11V: Tetanus

Case History: A previously well farmer was brought to the emergency room with a 4 hour history of rigidity of the muscles of the face, neck and trunk.

Alternative Diagnoses: Parkinson's disease, infection of the CNS, environmental chemical exposure, stroke

Strong Features: sustained contraction of facial muscles

Weak Features: intact alertness and mentation, opisthotonos, profuse sweating

Control Features: jaundice, periorbital edema

12V: Systemic Lupus Erythematosis

Case History: A pregnant lady presented with fever and joint pain

Alternative Diagnoses: rheumatic fever, flu, infection of joints, scarlet fever

Strong Features: malar rash,

Weak Features: discoid rash, parotid gland enlargement, anemia

Control Features: epicanthal folds, ptosis

13V: Cushing's Disease

Case History: A woman presented with generalized muscle weakness and amenorrhea

Alternative Diagnoses: pregnancy, thyroid problems, pituitary disorder, polycystic ovaries

Strong Features: round face, buffalo hump

Weak Features: hirsutism, hyperpigmentation

Control Features: supraclavicular nodes, micromandible

16V: Mumps

Case History: A boy presented with fever, malaise and pain on swallowing

Alternative Diagnoses: eppiglottitis, viral pharyngitis, strep throat, tonsillitis

Strong Features: inflammation and swelling of the parotid glands

Weak Features: lymphadenopathy, ears displaced up- or outward, edema

Control Features: coarse features, pigmented nevi

Appendix A (page 3)

17V: **Hyperthyroidism**

Case History: A woman presented with weight loss and increased frequency of bowel movements

Alternative Diagnoses: bowel cancer, malabsorbtion problem, irritable bowel disease, Crhon's disease

Strong Features: goiter, exopthalmos

Weak Features: widening of palpebral fissure, flushed skin

Control Features: facial weakness, spider angioma

18V: **Hypothyroidism**

Case History: A woman presented with increased fatigue and weakness

Alternative Diagnoses: Cushing's disease, Guillan-Barre, cancer, depression

Strong Features: myxedema, enlarged protruding tongue

Weak Features: sparse hair, pale skin

Control Features: supraclavicular nodes, low set ears

19V: **Pancreatitis**

Case History: A man came to the emergency room with the complaint of sudden onset of severe right sided abdominal pain which radiated through to the back 12 hours following a large meal

Alternative Diagnoses: biliary cholic, hepatitis, appendicitis, cholecystitis

Strong Features: jaundice, abdominal distention

Weak Features: profound weight loss, erythematous skin nodules due to subcutaneous fat necrosis

Control Features: ptosis, fever

21V: **Cirrhosis of the liver**

Case History: A man presented with weakness, weight loss, vomiting, and epigastric discomfort

Alternative Diagnoses: duodenal ulcer, gastritis, esophageal cancer, gastric carcinoma

Strong Features: wasting, spider angioma

Weak Features: jaundice

Control Features: webbed neck, narrow maxilla, edema

Appendix A (page 4)

22V: Congestive Heart Failure

Case History: A man was brought to the emergency department with shortness of breath

Alternative Diagnoses: COPD, pulmonary edema, pneumonia, myocardial infarction

Strong Features: chronic ill appearance, engorged neck vein

Weak Features: orthopnea, azotemia

Control Features: mental retardation, low set ears

23V: Nephrotic Syndrome

Case History: A girl presented with anorexia, malaise and bubbly frothy urine

Alternative Diagnoses: urinary tract infection, diabetes, infection, rheumatic fever

Strong Features: ill appearance, periorbital edema, puffy face

Weak Features:

Control Features: pallor, pigmented nevi, exopthalmos

25V: Leukemia

Case History: A girl presented with flu-like symptoms, epistaxis and a sore throat

Alternative Diagnoses: upper respiratory tract infection, flu, pharyngitis, mononucleosis

Strong Features: anemia, fatigue

Weak Features: petichiae, lymphadenopathy

Control Features: jaundice, muscle wasting

27V: Acute Glomerulonephritis

Case History: A child was brought to the emergency room because he was passing dark colored urine

Alternative Diagnoses: dehydration, something eaten, trauma (abuse), hemolysis

Strong Features: fever, mild anemia, periorbital edema

Weak Features: azotemia

Control Features: enlarged scalene nodes, low set ears

Appendix B: Sample of the 1 alternative condition

3V. A man came to the emergency room with haematemesis and mild epigastric pain.

A) This is being evaluated as a possible case of ***stomach cancer****. Please look at the photograph, and identify all clinically important features.*

_______________ _______________ _______________

_______________ _______________ _______________

_______________ _______________ _______________

_______________ _______________ _______________

B) Considering all the data, please rate the likelihood of ***stomach cancer****.*

|____1___|____2___|____3___|____4___|___5_____|___6_____|___7_____
|

Highly	Unlikely	Somewhat	Uncertain	Somewhat	Likely	Highly
Unlikely		Unlikely		Likely		Likely

Appendix C: Sample questions for the feature rating task

3V. Does this person have the following features?

-jaundice

1	2	3	4	5
very unlikely	somewhat unlikely	uncertain	somewhat likely	very likely

-epicanthal folds

1	2	3	4	5
very unlikely	somewhat unlikely	uncertain	somewhat likely	very likely

-weight loss

1	2	3	4	5
very unlikely	somewhat unlikely	uncertain	somewhat likely	very likely

-supraclavicular nodes

1	2	3	4	5
very unlikely	somewhat unlikely	uncertain	somewhat likely	very likely

-enlarged scalene nodes

1	2	3	4	5
very unlikely	somewhat unlikely	uncertain	somewhat likely	very likely

-myxedema

1	2	3	4	5
very unlikely	somewhat unlikely	uncertain	somewhat likely	very likely

Appendix D: Sample of the 5 alternative condition

3V. A man came to the emergency room with haematemesis and mild epigastric pain.

A) This history may suggest multiple diagnostic possibilities, including ***liver cancer, stomach cancer, gastric ulcers, esophageal varices,*** *and* ***lung cancer****. Please look at the photograph, and identify all clinically important features.*

________________ ________________ ________________

________________ ________________ ________________

________________ ________________ ________________

________________ ________________ ________________

B) What additional diagnoses might you consider?

____

____

____

C) Considering all the data, please rate the likelihood of the following diagnoses.

-liver cancer

I____1___I____2___I____3___I____4___I___5____I___6____I___7____
I

Highly Unlikely	Unlikely	Somewhat Unlikely	Uncertain	Somewhat Likely	Likely	Highly Likely

-stomach cancer

|____1___|____2___|____3___|____4___|___5____|___6____|___7____
|

Highly Unlikely	Unlikely	Somewhat Unlikely	Uncertain	Somewhat Likely	Likely	Highly Likely

-gastric ulcers

|____1___|____2___|____3___|____4___|___5____|___6____|___7____
|

Highly Unlikely	Unlikely	Somewhat Unlikely	Uncertain	Somewhat Likely	Likely	Highly Likely

-esophageal varices

|____1___|____2___|____3___|____4___|___5____|___6____|___7____
|

Highly Unlikely	Unlikely	Somewhat Unlikely	Uncertain	Somewhat Likely	Likely	Highly Likely

-lung cancer

|____1___|____2___|____3___|____4___|___5____|___6____|___7____
|

Highly Unlikely	Unlikely	Somewhat Unlikely	Uncertain	Somewhat Likely	Likely	Highly Likely

Appendix E: Sample of the GEN condition

3V. A man came to the emergency room with haematemesis and mild epigastric pain.

A) Please look at the photograph. Based on it, and on the information received above, please list the most likely diagnoses. Also, please rate the likelihood of each diagnosis.

1.________________

I____1___I____2___I____3___I____4___I____5____I___6____I___7____
I

Highly Unlikely	Unlikely	Somewhat Unlikely	Uncertain	Somewhat Likely	Likely	Highly Likely

2.________________

I____1___I____2___I____3___I____4___I____5____I___6____I___7____
I

Highly Unlikely	Unlikely	Somewhat Unlikely	Uncertain	Somewhat Likely	Likely	Highly Likely

3.________________

I____1___I____2___I____3___I____4___I____5____I___6____I___7____
I

Highly Unlikely	Unlikely	Somewhat Unlikely	Uncertain	Somewhat Likely	Likely	Highly Likely

B) Now, please look at the photograph, and list all clinically important features.

________________ ________________ ________________

________________ ________________ ________________

Appendix F: Sample of the ND condition

3V.

A) *Please look at the photograph, and identify all clinically important features.*

____________________ ____________________ ____________________

____________________ ____________________ ____________________

____________________ ____________________ ____________________

____________________ ____________________ ____________________

Appendix G
General knowledge study: Sample questions.

How confident are you (0 - 100%) that the name of the avenue that immediately follows Atlantic Avenue in the game of Monopoly is:

a. Ventnor ____________

b. Something else ____________

Plausible alternative:

-Marvin Gardens

-Illinois

Unlikely alternatives:

-Park Place

-Baltic

Not in Domain alternative:

-King

-QEW

Appendix G (page 2)

How confident are you (0 - 100%) that the capital of Australia is:

a. Canberra ____________

b. Something else ____________

Plausible alternative:

-Sydney

-Melbourne

Unlikely alternatives:

-Alice Springs

-Perth

Not in Domain alternative:

-Hong Kong

-New Guinea