INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

THE INFLUENCE OF DIFFERENTIALLY PROCESSING EVIDENCE ON DIAGNOSTIC DECISION-MAKING

By

KEVIN WAYNE EVA, B. SC. (HONOURS)

A Thesis

Submitted to the School of Graduate Studies

In Partial Fulfillment of the Requirements

For the Degree

Doctor of Philosophy

McMaster University

© Copyright by Kevin Wayne Eva, 2001

BIASING ATTENTION DURING DIAGNOSTIC DECISION-MAKING

DOCTOR OF PHILOSOPHY (2001) **McMaster University**
(Psychology) **Hamilton, Ontario**

TITLE: **The influence of differentially processing evidence on diagnostic decision-making**

AUTHOR: **Kevin Wayne Eva**
B. Sc. (Honours) (McMaster University)

SUPERVISOR: **Dr. Lee R. Brooks**

NUMBER OF PAGES: **x, 115**

ABSTRACT

By definition, the act of decision-making requires the consideration of multiple alternatives. Medical decision-making is no exception in that to decide whether or not a particular diagnosis is probably correct, a clinician must consider other diagnostic possibilities. Failure to do so is referred to as premature closure. If clinical instruction is provided regarding this diagnostic hazard, it tends to consist of advice to think of differential diagnoses before concluding one's diagnostic search. However, it is not yet known if such instruction is effective, let alone sufficient to eliminate diagnostic errors that arise from the tendency to under-weight non-focal hypotheses.

I examined this issue by presenting case histories to experimental participants and manipulating the diagnoses that were explicitly presented with a request to assign probability ratings, a paradigm extensively used by Tversky and Koehler (1994). Fluctuations in the probability assessments provided an empirical measure of the degree to which diagnostic hypotheses were considered when they remained implicit in an "all other diagnoses" category. Results showed that diagnosticians tend to under-weight diagnostic alternatives that they are not explicitly asked to evaluate (Experiments 1 & 2). Furthermore, this effect does not solely arise from non-generation of alternative diagnoses. Diagnoses that medical students themselves generate are under-weighted relative to when the same diagnoses are presented explicitly (Experiment 3). By

extension, this suggests that educational instruction to 'consider alternatives' is insufficient to help students eliminate premature closure.

I hypothesized further that premature closure might result from a confirmation bias because the explicit presentation of a diagnostic alternative might cause diagnosticians to differentially process the evidence relevant to the diagnostic possibilities. That is, judges might focus their attention on the signs and symptoms of a case that are consistent with the diagnosis they are asked to evaluate at the expense of reducing consideration of the information consistent with other hypotheses. The results of Experiments 4 to 6 support this hypothesis. In these experiments it was found that participants who were prompted to re-evaluate the evidence were less susceptible to the biasing influence of the explicitly presented alternatives as long as this re-evaluation took place at the moment of decision-making. This was true even though no additional information was provided through prompting.

Finally, additional factors that might lead to biased processing of the available evidence were examined. Formal medical terminology will cause differential processing, possibly because it is more mnemonically effective (Experiment 7). Diagnosticians are less susceptible to under-weighting non-focal alternatives when they themselves generate the focal hypotheses (Experiment 8). And finally, the diversion of attention created by the explicit mention of a specific diagnosis is not driven purely by an analytic shift of attention towards that particular diagnosis. Non-analytic influences are also influential (Experiment 9). The importance of these results for medical education and psychological theory of subjective probability is discussed.

ACKNOWLEDGEMENTS

I have been fortunate through much of my life to have the opportunity to learn from some truly exceptional role models. My graduate career has been no different and I am greatly indebted to all three members of my supervisory committee: Lee Brooks, Geoff Norman, and Bruce Milliken. All three have invested a great deal of time and energy for which I am more grateful than they could possibly realize. It goes without saying that any skills I have as a scientist have been greatly enriched by my relationship with all three men, but I am especially grateful that their tutelage and friendship extends beyond academics. Lee, thanks for teaching me to count cards. Geoff, thanks for the woodworking tips. And Bruce, my ability to step back into an open jumper is greatly improved thanks to you. I would also like to express thanks to two other faculty members from our department. Martin Daly and Margo Wilson opened my eyes to the energy and excitement that can be enjoyed in the world of academia and for that I will always be grateful. To Rose Hatala, Alan Neville, and John Cunnington, thanks for taking the time to share your clinical expertise. This dissertation could not have been completed without your generous contributions.

I am very appreciative of the close friends that I made during my time in the department. Tim Wood, thanks for repeatedly getting me into trouble and then right back out again. Thanks also to you and Jim Debner for the camaraderie around the dartboard;

mathematical modeling has never been so much fun. To Karmen Bleile, thanks for interrupting me every day. I always looked forward to your visits and will continue to do so. Also, to Jason Tangen, my co-movie night producer, and Melanie McKenzie, thanks for sharing so many interests, including the trips to Café 2000. A special thanks goes to Noel Jones and Suzanne Brown who opened their house to Betsy and I and who have been very supportive in all of our endeavours. It has been a pleasure watching Noah grow and I look forward to doing the same with Owen.

I would also like to thank family, both my own and Betsy's, for their unrivaled encouragement, confidence, and patience. This endeavour was made easier thanks to the strong support network that was always available.

Finally, I would like to dedicate this thesis to the person who has made the last quarter of my life the best yet. Betsy Agar, thanks for making me laugh, for helping me keep sight of the important things in life, and for motivating me to improve each and every day.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1

Introduction

Performing complex categorization tasks such as those faced by diagnosticians often requires the generation and evaluation of multiple diagnostic alternatives. Imagine that a diagnostician is asked to comment on a diagnosis that has been proposed by a colleague. Clearly, to decide whether or not that diagnosis is probably correct, other diagnostic possibilities must be considered. That clinical instructors are aware of this assertion is evident every time a student is admonished to think of differential diagnoses. Failure to adequately consider alternative diagnoses is referred to as premature closure and is commonly viewed as one of the primary hazards that students must learn to overcome (Voytovich, Rippey, & Suffredini, 1985; Krenz, Wolf, and Stahl, 2000).

Despite the emphasis placed on avoiding premature closure, it is not known if instruction to think of alternative diagnoses is sufficient to eliminate these types of diagnostic errors. The prevalence of confirmation biases that have been identified in the psychology literature (see Reisberg, 1997 for an introduction) suggests that evaluation of a clinical case with a particular diagnosis in mind might cause that diagnosis to assume priority over other diagnoses. That is, assessing a clinical case in the context of a proposed diagnosis might lead a judge to focus his or her attention on the evidence within the case that is consistent with the specific diagnosis being appraised.

Medical personnel may be particularly susceptible to confirmation biases as the data that allow diagnostic decisions to be made are inherently ambiguous (Brooks, Norman, & LeBlanc, 2000; Hatala, Norman, & Brooks, 1999). A single cluster of symptoms can often

be construed as supportive of two or more diagnoses. Furthermore, clinically relevant symptoms are easily confused with normal variability in the appearance of features.

The series of studies described in this dissertation were conducted to examine the effect of such ambiguities on premature closure. More specifically, Experiments 1 to 3 illustrate that an instruction to 'consider alternatives' is unlikely to help students eliminate premature closure, because the consideration paid to self-generated diagnoses tends to be less than that paid when the same diagnoses are explicitly presented. Experiments 4 to 7 explore the mechanism that produces this under-weighting. These experiments concurrently test whether or not instructing students to "re-evaluate the evidence while making the final judgment" might reduce this bias. Experiment 8 tests whether individuals are less susceptible to under-weighting non-focal diagnoses in the absence of an explicit diagnostic suggestion relative to when one is presented. Finally, Experiment 9 examines non-analytic influences on probability judgments. The arguments relevant to each study are elaborated upon in the appropriate chapters following a more general discussion of the research paradigm used.

Subjective Probability

From a purely normative standpoint, judgments about the likelihood of an event should not be influenced by the way in which the question is asked. However, if the study of subjective probability has revealed anything about human tendencies when making these types of decisions, it is that we often fail to follow the rules inherent to standard probability theory (e.g., Tversky and Kahneman, 1983). The way in which a question is asked can have a profound influence on the response received. For example, the probability that "a randomly selected person will die of natural causes" is usually judged as lower than the probability that "a randomly selected person will die of cancer, lung disease, or any other

natural cause." The set-theoretic extension (i.e., dying of a natural cause) is identical in both the former implicit disjunction and the latter explicit disjunction and as a result, logic dictates that the probability of both events is the same. Although the judge knows that both cancer and lung disease are examples of natural causes of death, posing the question in the more explicit manner tends to increase the probability rating assigned to the event.

This phenomenon, called "the unpacking effect" by Tversky and Koehler (1994), is extremely robust. Pioneering work in this area, conducted by Fischoff, Slovic, and Lichtenstein (1978), revealed that splitting a hypothesis into its specific components increases the probability ratings assigned by both mechanics and laypeople when they are asked to assess the cause of a car's failure to start. For example, mechanics rated the probability that "the cause of failure is something other than the battery, the fuel system, or the engine" as being greater when more specific causes such as the steering system or the ignition system were mentioned explicitly relative to when these causes remained implicit. Since 1978 the unpacking effect has been observed when lay-people evaluate the probabilities associated with (1) possible causes of death, (2) college majors (Koehler, Brenner, and Tversky, 1997), (3) suspiciousness of characters in crime stories (Teigen, 1982), (4) the outcomes of sporting events, (5) changes in the Dow-Jones Industrial Average, and (6) future outside temperature (Tversky and Fox, 1995). It occurs whether judgments are requested in the form of probabilities, frequency assessments (Tversky and Koehler, 1994), or confidence ratings (McKenzie, 1998). Similarly, it has been shown that individuals offered health insurance that covers hospitalization for any disease or accident are willing to pay more than those who are offered health insurance that covers hospitalization for any reason (Johnson, Hershey, Meszaros, and Kunreuther, 1993).

Furthermore, the unpacking effect is not simply a property of naïve judgments. It is observed when expert physicians evaluate the probability of medical diagnoses as they relate to clinical cases (Redelmeier, Koehler, Liberman, and Tversky, 1994) and when professional

bookmakers generate the odds for sporting events (Ayton, 1997). Similarly, this phenomenon can be observed when people make judgments about their personal lives (Mulford & Dawes, 1999). The unpacking effect would be expected if subjects assume that the hypotheses mentioned were presented due to their being probable. However, probability judgments reveal the unpacking effect even when researchers are careful to remove such prestige from the suggested alternatives (Tversky and Koehler, 1994). This list of examples is not comprehensive, but each example supports Tversky and Koehler's (1994) claim that individuals reliably under-weight the strength of support maintained by individual components of an implicit disjunction. As such the unpacking effect may be a useful tool for measuring the extent to which implicit alternatives have been considered.

To model the unpacking effect, Tversky and Koehler (1994) have proposed Support Theory, a non-extensional account of subjective probability. The studies reported in this thesis neither reinforce nor refute the fundamental axioms of Support Theory. Nevertheless, the theory is described here as it provides an orienting framework within which many of the findings can be interpreted.

Support theory has the very promising characteristic of being based on descriptions of events rather than extensions of events as in standard probability theory. According to Support Theory, each hypothesis entertained by a person is accompanied by a degree of support – a personal assessment of the strength of the evidence in favour of that hypothesis. Judged probability, by analogy to standard probability theory, is then calculated as a ratio of support in favour of the hypothesis to the sum of support for the hypothesis and support for all alternatives:

$$P(A, B) = \frac{s(A)}{s(A) + s(B)} \qquad (1)$$

where P (A, B) = probability of 'A' rather than 'B' (assuming that either hypothesis 'A' or hypothesis 'B' obtains) and s (A) = support for A.

Critical to the applications of Support Theory is the subadditivity assumption, namely that expressing an hypothesis such as "dying of natural causes" as an explicit disjunction will generally produce a greater total amount of support than when the component hypotheses remain implicit. In turn, rating each of the terms of the explicit disjunction independently tends to result in an even greater amount of support and hence an even higher judged probability (Rottenstreich and Tversky, 1997). For example, when each component is evaluated independently, the summed judged probability that a randomly selected person will die of "lung disease," "cancer," or "some other natural cause" will at least equal, but will usually exceed the support for the judged probability when these components are evaluated as a whole:

$$s(A) \leq s(A_1 \vee A_2 \vee A_3) \leq s(A_1) + s(A_2) + s(A_3) \qquad (2)$$

where $(A_1 \vee A_2 \vee A_3)$ is an explicit disjunction of the implicit disjunction 'A' and $s(A_1) + s(A_2) + s(A_3)$ are the sum of support ratings for the component hypotheses A_1, A_2, and A_3, evaluated independently. Using the above example, 'A' = death due to natural causes, $(A_1 \vee A_2 \vee A_3)$ = cancer, lung disease, or some other natural cause of death.

It follows from these two formulations that the probability assigned to the focal diagnosis (A) should decrease when potential alternative diagnoses are mentioned explicitly, thereby unpacking the residual category (B). Continuing the example, equation 2 suggests that explicitly stating that people might die in traffic accidents or by drowning should increase the support held in favour of dying of unnatural causes. Such unpacking would

then be expected to increase the support for B, thereby raising the denominator of the right side of equation 1 and reducing P (A, B).

Tversky and Koehler (1994) speculated that there might be two psychological bases for this subadditivity phenomenon. First, they argued that unpacking an hypothesis into component hypotheses might remind people of possibilities they may have overlooked. Second, the explicit mention of a component hypothesis could potentially increase the attention one assigns to it (salience), thereby increasing its perceived support.

Using Subjective Probability to Study Premature Closure

One issue not addressed by Tversky and Koehler's formulation of subjective probability is the role of the prevalence/plausibility of the unpacked components. This variable is potentially important, however, when trying to understand the problem of premature closure. As such, the prevalence/plausibility of unpacked components serves as the focus of Experiments 1 and 2. In particular, one might predict that the explicit mention of highly probable differential diagnoses will have less effect on the probability rating assigned to a focal diagnosis than the explicit mention of less common differential diagnoses. Medical personnel would presumably have already generated highly probable diagnoses while evaluating the clinical case. In the absence of their explicit presentation, highly probable alternatives are the differential diagnoses that should come to mind most readily, and once brought to mind, they should be evaluated with great care. Therefore, the change in processing that occurs upon the explicit presentation of these alternatives should be small relative to the change that occurs upon the explicit presentation of more obscure diagnoses.

In contrast, if diagnosticians either do not generate even the most likely differential diagnoses or do not consider those alternatives to the extent that they warrant when

presented explicitly, then the magnitude of the unpacking effect should be proportional to the probability of the alternatives that are unpacked. That is, unpacking highly probable alternatives should lead to a greater decrease in the probability assigned to a focal diagnosis than unpacking less common differential diagnoses. This pattern of responses would be more disconcerting.

Relevant to this issue is a model developed by Brenner and Koehler (1999) that extended Support Theory's formulation of subadditivity. Among other things, they asked subjects to make probability judgments regarding who was likely to win the 1998 Academy Award for Best Actor. Using these judgments they modeled the extent to which support for component hypotheses in the residual category are discounted (i.e., under-appreciated). Modeling these data by assigning each hypothesis a particular weight (rather than assigning a single 'global' weight to the entire residual hypothesis), Brenner and Koehler reported that each component of the residual hypothesis appears to be discounted to some extent. In addition, they found greater variability for high-support (strong) hypotheses relative to low-support (weak) hypotheses in the amount of subjective weight assigned to each hypothesis as a function of whether the hypothesis was focal or part of the residual. Specifically, stronger components of the residual tended to be discounted to a greater extent than did weaker components. Consistent with this result would be the latter of the two potential patterns outlined above – a greater decrease in the probability assigned to the focal diagnosis as the strength of the unpacked (non-focal) alternative diagnoses increases. That is, the contrast between the amount of support perceived in the packed and unpacked versions of the residual category (s(B)) should be greater when stronger (more probable) alternatives are unpacked relative to when weaker (less common) alternatives are unpacked.

While this model is relevant to the issues of current interest, it does not preclude the current work. As stressed by Brenner and Koehler, "it would be a mistake to interpret the local weight associated with a component hypothesis as a direct measure of its salience or

contribution to the support assigned to the residual hypothesis as a whole" (1999, p.45). In other words, although Brenner and Koehler's model appears to suggest that unpacking highly probable alternatives will have a larger influence on the probability assigned to a focal diagnosis than unpacking less probable alternatives, their model alone does not allow that assumption. Whether or not a particular component of the residual hypothesis is thought of will affect the degree to which that component is discounted. Brenner and Koehler's model assumes that the probability of a focal diagnosis is assessed against a composite residual hypothesis (i.e., without assessment of particular components). In manipulating the *a priori* probability of the non-focal diagnoses, I intuit that the likelihood with which diagnosticians will have considered the specific components of the residual category that are unpacked will be altered. Therefore, it is inappropriate to use the properties of local weights outlined by Brenner and Koehler to deduce the pattern of probabilities that will be assigned by diagnosticians.

Furthermore, Brenner and Koehler's model may not be directly applicable in the current context since there are two important differences between the decisions that are made by diagnosticians and those made by the participants in Brenner and Koehler's behavioural studies. As mentioned, Brenner and Koehler had participants make probability judgments regarding the 1998 Academy Award for Best Actor. They also requested probability judgments regarding the Best Picture Oscar, the 1998 NCAA Men's Basketball Championship as well as frequency estimates regarding the number of students that were expected to be majoring in one of four possible social science programs. At the time of the study a limited number of 4 or 5 potential alternatives existed in each case, each alternative being plausible and each one known to the judge. For example, in the latter case, the experiment began by informing participants that the social sciences program at a large Midwestern state university consisted of four different majors: Economics, sociology, political science, and psychology. The set of diagnoses at the disposal of medical personnel

is much greater. In addition, when raised with reference to a particular medical case history, diagnoses exist for which the perceived probability will be more variable than any of the alternatives in Brenner and Koehler's studies.

These distinctions are important given the two hypothetical patterns of data outlined earlier. On one hand, the vast number of diagnoses that are potentially available to diagnosticians allows the possibility that surprise might be registered in response to the unpacking of a less common diagnosis. Such surprise might increase the amount of attention paid to non-focal alternative diagnoses, thereby resulting in a larger unpacking effect when less probable diagnoses are unpacked. In contrast, the opposite effect might occur because diagnosticians maintain an expertise to which undergraduate psychology students are not privileged. If this expertise allows judges to recognize when particular diagnoses do not warrant much attention, then improbable diagnoses should have little effect on the consideration of the case history. As a result, the unpacking effect would be expected to be smaller when improbable diagnoses are unpacked.

To this point, consideration of the effect of unpacking the residual category's component hypotheses has focused primarily on the influence of varying the probability of the diagnoses that are unpacked. A related issue that may also be relevant, however, is the influence of the plausibility of the diagnoses that are unpacked. That is, a particular diagnosis might have a relatively high prevalence within the population (i.e., highly probable), but remain an implausible diagnosis within the context of a specific case history. If the explicit presentation of non-focal implausible diagnoses results in an unpacking effect, then the robustness of this phenomenon will be confirmed in a different manner than described previously. Another possibility, however, is that unpacking implausible alternatives might result in their simply being ignored, which might in turn have little influence on the probability rating assigned to the focal diagnosis. In fact, if implausible diagnoses are unpacked, then it seems reasonable that these alternatives will be dismissed

immediately. The act of ruling out a diagnosis might in turn decrease the support one perceives for the residual category, thereby increasing the probability assigned to a focal diagnosis.

To my knowledge such superadditivity has not yet been observed. However, previous studies, by design, have unpacked only plausible non-focal alternatives. Furthermore, although Redelmeier et al. (1995) have reported an unpacking effect produced by expert physicians, most work in this area has used undergraduate psychology students. Although these subjects undoubtedly had some knowledge of the domains they were asked to evaluate (e.g., causes of death), they are unlikely to have had much experience with these types of judgments and do not maintain enough expertise to recognize when a particular alternative hypothesis is implausible.

Chapter 2 describes a pair of experiments that were conducted to determine whether the plausibility/probability of the non-focal alternatives will influence the probability assigned to a focal diagnosis. Subjects were diagnosticians in medicine (clinical clerks) and automotive technology (3 levels of expertise). These are individuals who have been trained to maintain an active role in evaluating clinical cases and are constantly placed in the position of having to generate multiple diagnoses in addition to those that are explicitly provided. In addition, they are participants who should have sufficient expertise to recognize when a diagnostic alternative can be eliminated from consideration.

CHAPTER 2

The strength of alternatives effect

Experiment 1

Similar to Redelmeier, Koehler, Liberman and Tversky (1995), I view the medical field as a promising area to study subjective probability judgments. This domain is rich with problems that need to be solved by weighing the likelihood of potential diagnoses, and clinicians are trained from the very beginning of their medical education to avoid prematurely closing their investigation by generating alternative diagnoses. Redelmeier et al. showed that the estimated probability of a particular medical diagnosis was substantially higher when it was a member of a short list than when it was a member of a longer list presented to physicians. In addition, they demonstrated that this "unpacking effect" does not reflect a simple incapacity to deal with probabilities – participants' management decisions were also influenced by the alternatives explicitly presented as possibilities. For example, the diagnosis sinusitis does not typically require a CT scan, but other competing diagnoses do. When diagnosticians were asked to evaluate the probability that sinusitis was the correct diagnosis for a particular case history, fewer respondents recommended a CT scan than when a longer list of potential diagnoses was presented. Analogous to the change in probability estimations, these data suggest that individuals under-weighted the non-focal alternative diagnoses when they were not explicitly mentioned.

Like Redelmeier et al., I asked participants to rate the probability that a given case history was representative of explicitly presented diagnoses. In the unpacked condition (5A) five diagnostic alternatives and a residual "all other diagnoses" category followed a

short case history. Unlike Redelmeier et al., instead of presenting two diagnoses and a residual category in the packed condition (1A), I presented only one diagnosis and a residual. This single diagnosis is called the focal diagnosis, because it was the only alternative presented explicitly in all conditions. The critical difference, however, between this study and previous experiments was that three unpacked conditions were used rather than just one. Three conditions were used because there are two ways in which medical diagnoses can be viewed as improbable: (1) Diagnoses have variable base rates – some disorders affect fewer individuals than do others, and (2) The likelihood of a diagnosis is also conditional upon the signs and symptoms observed – a given diagnosis may be very prevalent, yet unlikely as a diagnostic option for a patient, given his or her case history. To capture both of these possibilities, the diagnostic alternatives were categorized as (a) Common (High base rate - Consistent with the case history), (b) Uncommon (Low base rate - Consistent with the case history), and (c) Implausible (Inconsistent given the case history).

The subadditivity hypothesis, central to Support Theory, predicts that the probability assigned to the residual ("all other diagnoses") category (R) in the 1A condition should be less than or equal to the sum of the probability assigned to the residual category and the 4 non-focal alternatives (A_1 through A_4) in the 5A condition (i.e., $P(R \mid 1A) \leq P(R \cup A_1 \cup A_2 \cup A_3 \cup A_4 \mid 5A)$). This should be so regardless of the probability of the diagnostic alternatives that are unpacked. Since the presence of the residual category makes each diagnosis list exhaustive, an alternative, yet more concise way of describing this prediction is that the probability assigned to the focal hypothesis (F) in the 1A condition should be greater than or equal to that assigned to the same focal hypothesis in the 5A condition (i.e., $P(F \mid 1A) \geq P(F \mid 5A)$). For the sake of simplicity, it is this latter version that will be presented throughout this dissertation. Whenever P(F | Packed Residual) > P(F | Unpacked Residual) it will be argued that the component hypotheses that were unpacked

were under-weighted when the focal diagnosis was evaluated in isolation (i.e., the packed condition).

Method

Participants

The participant pool for this study was final year medical students (clinical clerks) from McMaster University's graduating class of 1998. At the time of study, these students were undertaking their third year of medical education through on-site training at University affiliated teaching hospitals. In the majority of cases, participants were asked to participate via phone contact, and if willing, were delivered a questionnaire to complete on their own time with the instructions that they were to do so independently and without the use of any reference materials. Of the 63 students contacted and provided with a survey, 42 (67%) had returned it by the end of the data collection period. Participants were paid $10 and offered feedback upon return of the completed questionnaire.

Materials

Questionnaire booklets were composed of 20 short case histories, each followed by the question: "How likely are each of the following diagnoses?" To assign specific diagnoses to the appropriate conditions, four physicians, all experts in internal medicine, were asked to aid me in designing my materials. One physician wrote twenty fictional case histories, based on twenty focal diagnoses, each of which was sufficiently indeterminate that other diagnoses could be viewed as plausible. For example, a potential case of fibromyalgia (a syndrome characterized by chronic musculoskeletal pain occurring in a relatively defined pattern and associated with marked tenderness at "trigger points" and which may be associated with disordered sleep, stress, anxiety, or other psychological disorders) was presented as:

> A 45-year-old female presents complaining of debilitating fatigue that has lasted for 6 months. Nothing appears to alleviate it. It has been associated with a diffuse aching musculoskeletal discomfort.

Next, a list of diagnoses was generated and two other physicians used a 7-point scale to independently rate the prevalence of each diagnosis. Finally, for each case, a fourth physician, in conjunction with the first, assigned three groups of four alternative diagnoses to each case based on both the average prevalence ratings received and their own expertise. The diagnoses, as explained earlier, were organized into three categories: a) Common, b) Uncommon, and c) Implausible. In two cases additional diagnoses, not previously rated for prevalence, were added to complete the problem set.

Procedure

Each participant saw all twenty case histories, each presented with its associated focal diagnosis in one of four experimental conditions. In the One-Alternative (1A) condition, after the case history was described, the only diagnostic alternatives presented were the focal diagnosis and a residual ("all other diagnoses") category. The remaining three conditions were all Five-Alternative conditions in which the focal diagnosis was offered within a list of four alternative diagnoses taken from one of the Common, Uncommon, or Implausible categories ($5A_C$, $5A_U$, or $5A_I$, respectively). In these conditions, the focal diagnosis was randomly ordered within the list. Table 1 provides an example by presenting the hypothesis list in each condition that was used for the fibromyalgia case described above.

In viewing all twenty cases, each participant saw five cases from each experimental condition and was asked to assign a probability (0 - 100%) to each alternative presented. They were told that the diagnoses for these cases were mutually exclusive and that the inclusion of an "all other diagnoses" alternative meant that each diagnostic list was also exhaustive and that, therefore, the sum of the ratings should be 100%. To determine

whether exposure to the 5A condition might cause participants to spontaneously unpack the residual diagnosis in the 1A condition, thereby voiding my experimental manipulation, half of the questionnaires presented five 1A cases before fifteen 5A cases (First condition). The other half presented five 1A cases mixed among fifteen 5A cases (Mixed condition).

Insert Table 1 about here

Results

As a manipulation check, an analysis of the probabilities assigned to the non-focal alternatives was performed. It was found that the medical students indeed rated the alternatives provided in a manner consistent with the estimates provided by the expert internists. Table 2 shows that the alternatives we categorized as Implausible were considered to be least likely in the eyes of the students. Also as expected, the diagnoses categorized as Common received the highest probability estimates while those rated as Uncommon fell in the middle.

Insert Table 2 about here

A potential problem was that the mixing of 1A and 5A cases might have voided my experimental manipulations. My concern was that initial exposure to the cases presented in the 5A condition might cause participants to spontaneously unpack the residual diagnosis in the 1A condition to a greater extent than they would normally. If this were true, one would expect the probability assigned to the focal diagnosis in the Mixed condition to be lowered relative to the First condition ($P(Focal \mid 1A_{Mixed}) < P(Focal \mid 1A_{First})$. As can be seen in

Table 3, there was no trend in this direction ($F(1, 152) = .002$, $MSe = 276.863$, $p > 0.95$). The data from these conditions were, therefore, pooled together for the remaining analyses.

If probability judgments are made independently for each alternative, the differences observed in Table 2 should have no effect on the estimates of the probability assigned to the focal diagnosis. This assertion becomes obvious when one realizes that all differential diagnoses are implicitly included in the residual "all other diagnoses" category when not presented explicitly.

However, as predicted by Support Theory, the presentation of these specified alternatives did have an effect. Upon unpacking the residual diagnosis into four additional diagnostic alternatives plus a new residual diagnosis, the sum of the ratings of these five options was observed to be greater than the average probability rating assigned to the residual diagnosis in the 1A condition. In other words, the focal diagnosis was rated as more likely in the 1A condition compared to when it was presented within a list of 4 other alternatives as shown in Table 3. Even when the specified alternatives were considered by experts to be Implausible given the case history, the predictions made by Support Theory held - the judged P(Focal | 1A) was found to be significantly greater than the judged P(Focal | $5A_I$), $t = 3.14$, 19 df, $p < 0.01$.

Insert Table 3 about here

A "strength of alternatives" effect was also found in that the likelihood of the alternatives had an impact on the probability assigned to the focal diagnosis. Within the three 5A conditions, the focal diagnosis was rated as most likely if the specified alternatives were Implausible and least likely if the specified alternatives were considered Common by the expert physicians (see Table 3), $F(3,152) = 6.78$, $MSe = 276.863$, $p < 0.001$. In other

words, the magnitude of the unpacking effect is directly related to the probability of the non-focal alternatives.

It is also possible to examine the data on a case-by-case basis using the probability assigned to the non-focal alternatives by the participants as an empirical measure of the strength of the alternatives. In doing so, each of the 20 cases provide three possible measures of the unpacking effect: $P(F \mid 1A)$ vs. $P(F \mid 5A_x)$ where x = C, U, or I. Analyzing these 60 data points yielded the same conclusion as outlined earlier. The judged probability of the diagnostic alternatives in the 5A conditions, when summed together ($P(A_1 \cup A_2 \cup A_3 \cup A_4 \mid 5A)$), was positively correlated with the difference between the estimated probability of the focal diagnosis in the 1A and 5A conditions ($P(F \mid 1A) - P(F \mid 5A)$), $r = 0.486$, 58 df, $p < 0.01$. So, using an empirically derived measure, the more probable the alternatives presented, the larger the effect of unpacking.

Of the 60 possible comparisons examined during this analysis, 10 failed to reveal change in the direction predicted by Support Theory (i.e., the judged P(Focal | 1A) was less than P(Focal | 5A)). All 10 of these cases fell in the Uncommon or Implausible categories. To further test whether highly unlikely alternatives were most likely to lead to an increase in the judged probability of the focal diagnosis upon unpacking the residual (i.e., opposite to the prediction of Support Theory), a chi-squared analysis was performed. Data were collapsed using the boundaries $P(F \mid 1A) - P(F \mid 5A) = 0$, and $P(A_1 \cup A_2 \cup A_3 \cup A_4 \mid 5A) = 30\%$. Note that 0 is the natural boundary for determining whether or not the predictions of Support Theory were confirmed and 30% was found to represent the median value when the probability estimated for each alternative was summed. Eight out of thirty (27%) cases in which $P(A_1 \cup A_2 \cup A_3 \cup A_4 \mid 5A) < 30\%$ failed to reveal subadditivity – a significantly higher portion than the 2/30 (7%) cases in which $P(A_1 \cup A_2 \cup A_3 \cup A_4 \mid 5A) > 30\%$. That is, significantly more of the 10 cases that showed a pattern different than that predicted by

Support Theory were presented with non-focal alternatives that fell in the lower end of the plausibility scale, $\chi^2 = 4.32$, 1 df, $p < 0.05$.

Discussion

I lend support to the model proposed by Tversky and Koehler (1994) by showing that on average, the probability rating of the focal hypotheses decreased when the residual category was unpacked. This was true even for those alternatives that were considered Implausible by expert raters. Also consistent with, but not necessarily predicted by Support Theory was the finding of the "strength of alternatives" effect. The magnitude of the unpacking effect is proportional to the probability of the alternatives explicitly presented. That is, contrary to a reasonable set of expectations, the decrease in the probability rating assigned to the focal diagnosis was greatest when the component hypotheses of the residual that were most likely to have been considered in the 1A condition were presented explicitly. This suggests that either participants did not generate even the most likely alternative diagnoses while working through the case or that consideration of the diagnoses that were self-generated was incomplete compared to when the same diagnoses were presented explicitly in the 5A condition.

It is interesting to note that 80% of the cases that did not follow the predictions of Support Theory were those that were presented with non-focal alternatives that participants considered to have very low probability. If this were a consistent finding, then situations in which very low probability alternatives were presented might constitute a fundamental violation of Support Theory. However, on average, the estimated probability of the focal hypotheses did decrease even when Implausible non-focal alternatives were presented. The low probability ratings assigned to the Implausible alternatives suggest that it might be

infeasible to find alternatives that are so Implausible that a consistent P(Focal) increase will be observed when the residual hypothesis is unpacked.

Potential problems with drawing the conclusion that superadditivity can not be observed on a consistent basis include (1) the level of expertise held by the medical students used as participants, and (2) the number of non-focal alternatives that were unpacked. First, as was argued in the introduction, I suspect that only experts would be able to adequately ignore diagnostic suggestions that are truly implausible. This might be especially true given the potential for diagnostic suggestions to be construed as arising from a prestigious source. The clinical clerks who participated in this study had a considerable amount of diagnostic experience, but perhaps they were not yet of sufficient expertise to ignore implausible alternatives entirely. Second, although the alternatives that were categorized as implausible were considered by the expert raters to be inconsistent with the case history provided, it is possible that, given a list of 4 diagnoses, one or two of them were not perceived as entirely implausible, thereby masking any superadditivity that might exist. The human body is a complex system and as a result, participants might have been hesitant to conclude that the probability of any specific diagnosis was zero. As plausible diagnoses tend to cause subadditivity, "Implausible" diagnoses that are perceived as plausible would counteract any superadditivity caused by the diagnoses that are perceived as implausible. Experiment 2 was conducted in an attempt to replicate Experiment 1 using (a) a different population of participants who have more experience within their domain of expertise, (b) fewer diagnostic alternatives, and (c) implausible diagnoses that seemed more obviously implausible.

Experiment 2

Experiment 2 was an attempt to replicate Experiment 1 with a new set of materials and a new participant population - motive technology students. In this Experiment, two rather than four diagnostic alternatives were unpacked for two reasons: (1) To test the robustness of the "strength of alternatives" effect, and (2) to ensure that all diagnoses that were labeled as implausible would truly be viewed as implausible.

Automotive mechanics were chosen as the study population for four reasons. First, they must undertake many activities analogous to those of physicians; in both professions, people are presented with symptomatic complaints relating to complex systems and must reason through examination and testing from symptom to solution. In addition, there is considerable variance in the frequency of both the types of problems observed by any individual and the duration of time between successive presentations of any given problem. Second, in both professions, probability estimates often act as tools for promoting understanding of the problem and potential actions to be undertaken. Third, the subject pool for motive technology students is larger than that of medical students, providing an analogous, yet easier to study population. Finally, the continuing education program in which these students were enrolled allowed us access to motive technology students of multiple levels, the most experienced of which had considerably more training than did the medical students who participated in Experiment 1. The Advanced group of students reported having had greater than 20,000 hours of work experience. That is, they had more than twice the number of hours that Ericsson and Charness (1999) suggest are required to become a true expert.

Method

Participants

The participant pool for this study was composed of mechanics in three different levels of the Motive Technology training program at Mohawk College. Most of the students had been working as automotive mechanic apprentices, and had returned to college for mandatory eight-week training courses. The Common Core (C.C.) class was nearing completion of the first such training session - these were the most novice students. The Basic Level (B) students were those who had previously taken the Common Core curriculum and were beginning to specialize in certain areas - they had been attending school one day a week for the year rather than full time for eight weeks. The Advanced Level students (A) had all of the in-school training of the C.C. and B groups and were nearing completion of the third and final level of formal training. In all, 26 Common Core students, 14 Basic Level students, and 25 Advanced Level students completed the questionnaire.

Materials

Questionnaire booklets were composed of 10 short case histories, each followed by the question: "How likely are each of the following diagnoses?" To assign specific diagnoses to a given category, symptoms were taken from a Motive Technology textbook (Duffy, 1994) along with a list of the possible causes of these symptoms as found within that text. The program manager of the motive technology program then read the case histories and their associated diagnoses and recorded estimates of the likelihood of each diagnostic alternative. The most likely diagnosis was named the focal Diagnosis. The next two highest were called Common Alternatives while the two alternatives with the lowest likelihood were named Uncommon Alternatives. In addition, two diagnoses were assigned randomly from the unused cases of completely separate systems and called Implausible

(e.g., in one case braking problems were used as diagnostic alternatives for a steering problem).

Procedure

Participants were asked to provide non-identifying demographic information including age and estimated number of lifetime hours spent working on cars so that I could get a better understanding of their experience. Afterwards, each participant saw all ten case histories, each one presented with its associated focal diagnosis in one of four experimental conditions. In the One-Alternative (1A) condition, after the case description was presented, the only diagnostic alternatives specified were the focal diagnosis and a residual "all other diagnoses" category. The remaining three conditions were all Three-Alternative conditions in which the focal diagnosis was offered within a list of two non-focal alternative diagnoses taken from the Common, Uncommon, or Implausible categories ($3A_C$, $3A_U$, or $3A_I$, respectively) and a residual category as described previously. Participants were asked to assign a probability (0 - 100%) to each alternative presented. They were told that the diagnoses for these cases were mutually exclusive and that the inclusion of an "all other diagnoses" alternative meant that each diagnostic list was also exhaustive and that, therefore, the sum of the ratings should be 100%. Three group sessions were run in which each of the three levels of students were asked to complete the questionnaires in their classroom.

Results

Two manipulation checks were performed. First, the number of hours experience working on automobiles increased in the expected direction (7,430 to 18,600 to 20,564 hours on average across the Common Core, Basic Level, and Advanced Levels, respectively). Second, once again, the probability ratings assigned to the non-focal diagnoses were greater in the Common condition than in the Uncommon condition and

these ratings were, in turn, greater than those assigned to the non-focal diagnoses in the Implausible condition for all three expertise levels. No significant differences were found across experience groups, $F(2,108) = 0.30$, $MSe = 414.187$, $p > 0.70$, so the data from these groups were pooled together for the following analyses.

The pattern of judged P(Focal) data from Experiment 2 replicated that from Experiment 1 across all three levels of expertise. Once again, the judged probability of the focal diagnosis was higher in the One-Alternative condition than in any of the Three-Alternative conditions including the 3A-Implausible condition ($P(\text{Focal} \mid 3A_I) < P(\text{Focal} \mid 1A)$), $t = 2.94$, 9 df, $p < 0.02$). Also, the "strength of alternatives" effect was present again. As can be seen in Table 4, for all three levels of expertise the higher the likelihood of the non-focal alternative diagnoses, the lower the judged probability of the focal diagnosis, $F(3,108) = 6.29$, $MSe = 414.187$, $p < 0.001$.

Insert Table 4 about here

Once again, looking at the data on a case-by-case basis, the summed probability of the unpacked diagnostic alternatives ($P(A_1 \cup A_2 \mid 3A)$), was positively correlated with the difference in the judged probability of the focal diagnosis across condition ($P(F \mid 1A) - P(F \mid 3A)$), $r = 0.381$, 89 df, $p < 0.01$.

Also, significantly more of the cases that did show a pattern opposite to the predictions of Support Theory again came from the lower end of the plausibility of alternatives scale. Using $P(A1 \cup A2 \mid 3A) = 30\%$ and $P(F \mid 1A) - P(F \mid 3A) = 0$ as the cutoffs once again, $\chi^2 = 4.99$, 1 df, $p < 0.05$.

Discussion

Whether three alternatives were presented to mechanics or five alternatives were presented to medical students, both groups of diagnosticians showed the "strength of alternatives" effect when making probability estimates. The similarities between mechanics and physicians have already been discussed, but it is also important to recognize the differences in the two fields. Automotive repair is often more structured in that it is possible to build trouble-shooting flow charts and follow a trial and error approach to repair when faced with more difficult situations. Physicians, on the other hand, must deal with systems that are often impossible to isolate and patients whose symptoms may be in a realm seemingly unrelated to their cause. There could also be a case made that the variation within both normal and abnormal is much higher in the human population than in the automotive population. Despite these differences, however, the "strength of alternatives" effect can be seen in both groups of students, suggesting that this finding is not just idiosyncratic to a particular domain of judgment, but rather, is a robust phenomenon that must be considered in the study of subjective probability.

The instances of superadditivity found in both experiments suggest that the probability assigned to a focal diagnosis is more likely to increase upon unpacking the residual category when the non-focal alternatives that are unpacked are implausible diagnoses. However, as in Experiment 1, the probability rating assigned to the focal diagnosis did decrease on average for all unpacked conditions.

Discussion of Experiments 1 & 2

As suggested by Tversky and Koehler (1994), the probability judgments made by the participants did not seem to be made simply on the basis of the likelihood of physical events as in standard probability theory. It appears that both the number and the likelihood

of alternatives that are explicitly presented heavily influence the probability ratings given by trained diagnosticians. I argue that the alternatives presented create a context within which the case and the diagnoses themselves are evaluated and that this context has the potential to alter the probability ratings assigned. The idea that a change in context (in this case, the presence or absence of diagnostic alternatives) might influence one's cognitive machinations (in this case, a diagnostician's belief as to the likelihood of a particular diagnosis) is neither outlandish nor novel. Studies of hindsight in both psychology (see Hawkins and Hastie, 1990 for a review) and medicine (Arkes et al., 1981) have revealed that people who know an event occurred tend to believe falsely that they would have predicted the reported event. I believe that mechanisms similar to those responsible for the hindsight effect are at play in the current paradigm. The presentation of a particular diagnosis might lead to differential processing of the evidence that is presented in the case history. Specifically, the explicit presentation of a particular diagnosis might cause diagnosticians to focus their attention on one or two features that are consistent with the diagnosis mentioned, thereby preventing them from appreciating the extent to which the evidence supports diagnostic alternatives that remain implicit. This hypothesis will be examined further in Chapters 4 and 5.

Regardless of whether this theoretical account of the unpacking effect is correct, it was somewhat surprising to note that the largest effect was found upon the explicit presentation of highly likely alternative diagnoses. Participants in both studies work in domains where they would be expected to have considered the common non-focal alternatives regardless of whether or not they were presented explicitly. Therefore, the current results appear to suggest that the unpacking effect was driven in large part by a change in salience rather than by simply not having considered probable alternatives. In other words, it appears that instruction to generate alternatives will be insufficient when trying to avoid premature closure as the attention paid to alternatives that are self-generated

is less than that paid to explicitly mentioned hypotheses. Of course, this argument assumes that all of the common diagnoses were generated when they remained implicit in the packed condition. Although these diagnoses are most likely to have been generated spontaneously, perhaps students self-generated only some subset of the common diagnoses and the remaining non-considered hypotheses caused the effects observed. This possibility will be examined further in Chapter 3.

Before examining this issue, however, one potential limitation on the generalizability of the "strength of alternatives" effect reported here should be considered. It is possible that the strength of alternatives effect might depend on the plausibility of the focal hypothesis. All of the diagnoses that were assigned the 'focal' label in these two studies were judged to be of relatively high probability. Support for this possibility has been presented by Koehler, Brenner, and Tversky (1997). Like Brenner and Koehler (1999), Koehler et al. modeled the extent to which support for component hypotheses in the residual category are discounted. Koehler et al., however, utilized global weights rather than local weights. Their model assumes that the greater the support for the focal hypothesis, the greater the extent to which component hypotheses in the residual are discounted. Therefore, greater subadditivity should be observed when a high-support hypothesis is pitted against its residual. Brenner and Koehler reported evidence in favour of this assertion. This does not, however, make the current results uninteresting even though the hypotheses used as focal diagnoses in the current pair of experiments were all strong (highly plausible) diagnoses. I have no reason to believe that the strength of the focal diagnosis would affect the qualitative nature of the strength of alternatives result, but Koehler et al.'s proposal suggests that the magnitude of the strength of alternatives effect might shrink if less probable focal diagnoses were used. The experimental design utilized does not allow me to directly test Koehler et al.'s assertion. Their model is consistent with my intuitions, however, that the amount of subadditivity might be influenced by the plausibility of the focal

diagnosis as diagnosticians are perhaps more likely to actively search for alternatives if presented with an implausible diagnostic suggestion. Furthermore, their model is consistent with my argument that support for a given hypothesis is dependent on the context created by the hypothesis list presented in the problem. Still, as I am mainly interested in studying subadditivity as it pertains to premature closure, the use of plausible focal diagnoses lends greater ecological validity. Therefore, the hypothesis that I believe warrants greater consideration is that diagnosticians might under-evaluate diagnoses that they themselves generate relative to when those same diagnoses are presented explicitly. This topic will be the focus of the next chapter.

CHAPTER 3

Under-weighting self-generated diagnoses

Experiment 3

The strength of alternatives effect outlined in Chapter 2 which showed a large unpacking effect when plausible alternatives were suggested might be interpreted as suggesting that participants under-appreciated diagnostic alternatives that they themselves generated relative to when the same alternatives were explicitly presented. However, the experimental design did not allow us to be certain that participants had actually considered the diagnoses that were rated as most likely. Five or three diagnoses were explicitly presented in the unpacked condition, thereby allowing the possibility that participants had not generated all of the plausible non-focal diagnoses that were unpacked while reading the case history in the packed condition.

The current study was designed to demonstrate the same result for alternatives that participants claim to have actually considered. Furthermore, I attempted to maximize the probability that a specific alternative diagnosis would come to mind even when not presented explicitly by using clinical cases previously shown by Cunnington, Turnbull, Regehr, Marriott, and Norman (1997) to be suggestive of two highly likely and roughly equiprobable diagnoses. Both manipulations should eliminate the unpacking effect if diagnosticians evaluate diagnostic possibilities that they themselves generate in the same way as they evaluate diagnoses that are explicitly provided.

A second goal of this study is to examine the ecological validity of any unpacking effect observed. Although subjective estimates of probability are believed to provide a valid

measure of participants' clinical decision making processes, it is possible that the act of assigning probabilities is a formal exercise that is not closely related to actual practice. So, the current study examined whether the unpacking effect extends beyond numerical estimates of probability. This issue was addressed by measuring whether or not patient management strategies, such as the requesting of diagnostic tests, are influenced by the explicit presentation of diagnostic alternatives. That is, if diagnosticians request more tests upon being presented two highly likely diagnostic alternatives relative to when they are presented just the focal diagnosis, I would have converging evidence that there is a tendency to under-weight alternatives that are not explicitly provided. Normatively, there should be no difference. Redelmeier et al. (1995) have previously shown that the likelihood that fourth year medical students will order a CT scan upon the presentation of a potential case of sinusitis is influenced by the number of alternative diagnoses that are explicitly mentioned. The current study attempts to further ensure the robustness and generalizability of their findings by using multiple cases and a more extreme manipulation.

Method

Participants

The participant pool for this study consisted of second year medical students from McMaster University's graduating class of 2001. A sample of tutorial leaders asked their students if they would participate. Those who agreed were run through the experiment in their tutorial groups during two sessions separated in time by an average of eight days (range = 4 to 14 days). Twenty students participated in four groups, but follow-up data could not be collected for one of the students, leaving 19 with a complete set of data. Upon completion of the second group session participants were paid $20 and provided feedback regarding both the clinical cases used and the purpose of the study.

Materials

Participants were presented 10 case histories, each of which was followed by one or two diagnostic hypotheses and a series of 5 questions. (1) Given the case history that you have just read, please assign a number between 0 and 100 indicating how likely you think it is that the case history is representative of the given diagnosis(es). In all conditions participants were told that the diagnoses were mutually exclusive and that the inclusion of an "all other diagnoses" alternative meant that each list was exhaustive, thereby indicating that the sum of the ratings assigned should be 100%. (2) Are there any diagnostic tests that you would like to see performed to aid you in your decision? If yes, please list them. (3) While reading the case history, did you consider any diagnosis apart from those listed above? If yes, please state the diagnosis that you consider to be the most likely differential. (4) Please rate your confidence (on a scale of 1 to 100) that you know the correct diagnosis. (5) Please rate the typicality of this case on a scale of 1 to 100. The latter two questions were intended to serve as dummy variables that would increase the likelihood that participants would not remember the exact probability assigned to any particular question.

Procedure

As mentioned, each of the ten cases have been shown to be suggestive of two diagnoses, both of which are highly likely and roughly equiprobable (Cunnington et al., 1997). One of each pair of diagnoses was randomly selected to be the focal diagnosis – the diagnostic alternative that would be presented with its associated case history across all conditions. In working through all 10 cases, each subject was shown five cases within each condition (i.e., Focal diagnosis alone vs. Focal + Alternative diagnosis), randomly mixed together. Approximately one week later each participant was shown the same 10 cases and asked to rate the original alternative(s) together with the alternative they had generated in response to question 3. If no alternative was generated, participants were simply shown the original alternatives a second time. Apart from adding the alternatives participants had

generated during the first pass the questionnaires used during the two sessions were identical.

Results

In completing all ten cases, 190 observations were generated that could be analyzed for the unpacking effect – a decrease in the probability assigned to a focal diagnosis upon the explicit presentation of additional diagnoses. Table 5 presents the average probability assigned to the focal diagnosis as a function of condition. First, a 2 (Session) x 2 (Number of alternatives presented during pass 1) x 2 (Diagnostic alternative: Generated versus Not generated) x 10 (Case) ANOVA was performed. A significant effect of 'number of alternatives' ($F(1, 156) = 12.365$, $MSe = 539.861$, $p < 0.01$) revealed that the probability assigned to the focal diagnosis was higher when presented in isolation than it was when presented in conjunction with a second diagnosis even though the alternative diagnosis was the most likely differential. An effect of 'diagnostic alternative' was also found ($F(1,156) = 6.009$, $MSe = 539.861$, $p < 0.02$), indicating that participants rated the focal diagnosis as more likely when they did not generate a plausible alternative diagnosis as indicated by their response to question 3. Case was the only other effect that reached significance ($F(9,156) = 4.746$, $MSe = 539.861$, $p < 0.01$).

To further demonstrate that the unpacking effect occurs even when the unpacked alternatives are diagnoses that participants have already considered, I performed a 2 (Session) x 2 (Number of alternatives presented during pass 1) x 10 (Case) ANOVA on only those observations in which a diagnostic alternative had been generated; that is, I used only the data presented in the third column of Table 5. The main effect of 'number of alternatives' persisted ($F(1,139) = 8.861$, $MSe = 521.250$, $p < 0.01$). In addition, a main effect of session was found ($F(1,139) = 16.375$, $MSe = 521.250$, $p < 0.01$) which indicates

that the probability assigned to the focal diagnosis was lower in Session 2 than in Session 1 even though the only difference between the two sessions was the explicit presentation during Session 2 of the diagnoses that participants claimed to have considered implicitly during Session 1. Case was, once again, the only other effect that achieved significance ($F(9,139) = 4.323$, $MSe = 521.250$, $p < 0.01$).

The effect of session was not observed when the same analysis was repeated for trials in which participants did not generate a diagnostic alternative (i.e., using only the data presented in column 4 of Table 5). This indicates that the effect was not simply a result of the passage of time. The number of observations in these cells was low, but examination of the means suggests that, if anything, the probability assigned to the focal diagnosis increased in Session 2 relative to Session 1 if no diagnostic alternative had been generated during Session 1 ($F(1,17) = 0.017$, $MSe = 392.034$, $p > 0.85$).

Insert Table 5 About Here

I also examined whether or not the phenomenon being illustrated by the probability ratings might influence patient management strategies by asking the participants to list the diagnostic tests that they would be interested in seeing performed. Participants requested more tests when two diagnoses were presented (mean = 3.464) relative to when the focal diagnosis was presented in isolation (mean = 2.989; $F(1,156) = 5.656$, $MSe = 2.155$, $p < 0.05$). This result suggests that the explicit presentation of diagnoses can influence the patient management strategies of diagnosticians in addition to altering their rating of another diagnosis' likelihood.

Finally, the effect of the number of alternatives presented on confidence ratings and typicality ratings were analyzed. No effect of session or 'number of alternatives' was found for either of these two variables.

Discussion

These results support the notion that individuals tend to under-appreciate self-generated diagnoses relative to diagnoses that are explicitly presented. Participants rated the originally presented diagnosis as less probable when the alternative they claimed to have considered implicitly was provided in a more explicit manner. That is, the unpacking effect was found even when the diagnostic alternative that was unpacked was one that the participants claimed to have considered while originally viewing the case. Furthermore, the fact that more tests were ordered when two diagnoses were presented relative to when just one diagnosis was presented suggests that under-weighting implicit diagnostic alternatives can alter patient management strategies as well as probability ratings.

Regardless of the mechanism responsible, these results indicate that the meaning of the verb 'to consider' should not be taken at face value. Having considered the plausibility of a diagnostic alternative can mean anything from having had the term come to mind to having performed a comprehensive analysis of the evidence for and against that particular diagnosis. Asking the participants to assign a probability rating to the diagnoses that they claim to have considered was sufficient to decrease the probability that they assigned to the focal diagnosis. This strongly suggests that the evidence in favour of the self-generated alternative was under-appreciated relative to when attention was focused on that alternative explicitly. In turn, this suggests that the problem of premature closure is not simply an inability to generate alternatives independent of a diagnostic suggestion. Rather, the explicit mention of a diagnostic possibility can reduce the degree to which self-generated diagnoses are considered.

That being said, the mechanism that causes individuals to under-weight alternatives that are not explicitly presented remains in question. As alluded to in Chapter 1, the effects observed might arise as a result of confirmation bias; the explicit presentation of a

diagnostic alternative might cause diagnosticians to differentially process the evidence relevant to the diagnostic possibilities. This could arise in at least three ways that are not necessarily exclusive of one another; the explicit presentation of a diagnostic hypothesis might influence the search for evidence, the interpretation of the evidence itself, and the construal of its relevance. Support for the plausibility of these hypotheses is widespread.

For example, it has been found that, when given the opportunity to select additional information (i.e., prevalence data), medical students (Kern & Doherty, 1982), residents (Wolf, Gruppen, and Billi, 1985) and physicians (Green & Yates, 1995) tend to seek data that is relevant to a single disease while ignoring information that is related to equally plausible differential diagnoses. This biased search for information need not be proactive in that it does not necessarily take place while the diagnostician gathers novel information. On the contrary, Anderson and Pichert (1978) have shown that memory-based retrieval of information is also influenced by the context within which the search takes place. When asked to recall information about a house, the type of information participants were able to remember was dependent on whether they had been asked to read the story from the perspective of a burglar or a home buyer. When subjects were later asked to adopt the opposite perspective, they were able to recall information that was not available to them during the first memory task. A plausible extension of this result is that the explicit presentation of a diagnosis might bias the memorial retrieval of features presented in the case history.

Furthermore, maintaining an initial focus on the diagnosis that is explicitly presented might make it difficult to realize that non-discriminating symptoms provide evidence for more than one diagnostic alternative. For example, the symptoms nausea and vomiting, in a case of an 18-year-old woman with right lower quadrant discomfort are supportive of both appendicitis and pelvic inflammatory disease. However, considering these symptoms as indicative of appendicitis might blind an individual to the possibility that these symptoms

can also be construed of as clinical manifestations of pelvic inflammatory disease. Brooks, LeBlanc, and Norman (2000) have provided evidence that supports this notion by reporting that the mere presentation of a diagnostic alternative can influence the interpretation of classic clinical features. Re-interpreting these features in light of self-generated diagnoses might prove to be difficult.

In summary, there is a great deal of evidence that ambiguities do exist in medical data that could allow diagnosticians to process clinical features differentially. If differential processing causes individuals to under-weight self-generated diagnoses, then eliminating differential processing should allow participants to avoid (or at least reduce) the bias illustrated by the unpacking effect. More specifically, adoption of a processing strategy that causes diagnosticians to re-evaluate the evidence while making their judgment should prevent under-weighting of diagnoses that remain implicit. Such a strategy should lead individuals to more fully utilize all of the evidence available to them. Three tests of this hypothesis will be described in Chapter 4.

CHAPTER 4

Reducing judgment bias by focusing attention on the evidence while a judgment is made

Experiment 4

A study that provides a useful lead for examining the effect of focusing attention on the evidence at the moment of decision was published by McKenzie (1998). He used a learning manipulation to illustrate that the way in which one learns to make diagnoses influences the degree to which one under-weights an alternative hypothesis. Undergraduate psychology students were taught to diagnose two fictional disorders named puneria and zymosis. During learning, each participant was presented with a series of patient profiles that consisted of the presence or absence of 16 (or 8 or 4 depending on the study) symptoms and they were asked to diagnose each patient. After making a judgment participants were given feedback consisting of "Correct" or "Incorrect" and they were told to adjust their categorization scheme accordingly until they were able to adequately diagnose both puneria and zymosis. The critical manipulation was whether students learned puneria and zymosis concurrently (Contrastive learning) or sequentially (Non-contrastive learning). Contrastive learners worked through a series of patients, each of which had either puneria or zymosis. Non-contrastive learners, on the other hand, worked through two sets of cases, the first of which were puneria or normal patients while the second were zymosis or normal patients.

In a series of four experiments McKenzie (1998) showed that students in the contrastive learning condition were more likely to take the alternative (non-focal) hypothesis into account when making their decisions. Compared with non-contrastive learners, contrastive learners were (a) less likely to select symptoms that supported both diagnoses equally when asked to identify symptoms that would allow them to distinguish between the two diagnoses, (b) better able to recognize that such symptoms did not provide additional information that would allow them to discriminate between the two diagnoses, and (c) more likely to assign frequency judgments, confidence ratings, and treatment decisions that were consistent with having considered the alternative. Because non-contrastive learners are unlikely to fully evaluate the non-focal diagnosis, this condition allows us to test the hypothesis that focusing attention on the evidence at the time of judgment might enable judges to make decisions in a less biased manner.

In his fourth experiment, McKenzie (1998) tested the hypothesis that differences in the tendency of contrastive and non-contrastive learners to under-weight non-focal diagnoses might be reduced by drawing non-contrastive learners' attention to the fact that there is an alternative diagnosis. This was done by mentioning the non-focal diagnosis within the question that was asked, a manipulation that can be thought of as analogous to a "think of differential diagnoses" instruction when trying to help someone avoid premature closure in a medical context. McKenzie had subjects assign a series of ratings to 16 test patients indicating the participants' confidence that the patient had a particular disorder. As all participants were told that every patient presented in the test phase had either puneria or zymosis, the sum of the average confidence ratings assigned by the group that was asked to assign a probability rating to puneria and the group that was asked to assign a probability rating to zymosis should equal 100. For example, patients who are very likely to have puneria should be assigned a high rating by the puneria-focal group and a low rating by the zymosis-focal group, thereby causing the sum of the average ratings assigned by both

groups to equal 100. The mean absolute deviation from 100, therefore, provides a measure of the extent to which the groups took the non-focal diagnosis into account when assigning their probability ratings. The further from zero, the greater the non-additivity, thereby illustrating a greater amount of bias.

One quarter of the subjects were asked to rate their confidence that puneria was the correct diagnosis while another quarter were asked to rate their confidence that zymosis was the correct diagnosis (Asymmetric Questions). Non-contrastive participants' responses revealed a mean absolute deviation from 100 of approximately 42, substantially different from the 0 that should be observed if both groups of respondents took the non-focal diagnosis into account when assigning their confidence ratings.

To test whether or not the degree to which non-contrastive learners under-weight non-focal diagnoses could be reduced by drawing their attention to the fact that there is an alternative diagnosis, McKenzie (1998) explicitly mentioned both diagnoses to the other half of the participants before asking them to assign confidence ratings to the same 16 test cases. That is, one quarter of the participants were asked to state how confident they were that each patient had puneria rather than zymosis while the final quarter were asked to rate their confidence that each patient had zymosis rather than puneria (Symmetric Questions). Asking non-contrastive learners to respond to a Symmetric question was intended to draw their attention to the presence of an alternative diagnosis. Indeed these participants' judgments showed less non-additivity than did those who were asked the Asymmetric question ($M = 27.5$). At the same time, the explicit mention of the non-focal diagnosis did not completely eliminate the bias. A mean absolute deviation from 100 of 27.5 is significantly less than 47.0, but remains substantially greater than 0, the point at which it could be said that no bias occurs.

In summary, McKenzie's results (1998) suggest that warning people of the presence of a diagnostic alternative reduces, but does not eliminate, the bias in diagnostic

judgments. Put in the current context, this result supports the view that instructing participants to consider alternative diagnoses is likely to be insufficient to eliminate premature closure. People appear to under-weight non-focal alternatives that do not elicit a judgment even when they are explicitly mentioned. This is especially surprising because there was only one alternative to the focal diagnosis in the McKenzie study. So, the effort required to consider the alternatives should have been minimal.

Of current interest is whether or not instructing diagnosticians to re-evaluate the evidence at the point of making a decision might lead judges to more fully consider the alternatives. As such, I replicated the noncontrastive condition in McKenzie's (1998) Experiment 4. In addition, a third group was tested. Participants in this group (the "Evidence" group) were prompted to focus on the evidence by a request to state which symptoms favoured puneria and which symptoms favoured zymosis before assigning their confidence ratings. Again, half of the participants in this group were presented with puneria as the focal diagnosis while the other half were presented with zymosis as the focal diagnosis. As with McKenzie's work, the outcome of interest was the mean absolute deviation from 100 after summing across the responses provided by these two groups.

Method

Participants

One hundred and twelve students enrolled in a first year Introductory Psychology course at McMaster University participated in this experiment for course credit. The stimuli were presented via computer projection to ten groups of 9 to 13 subjects (mean = 11.2).

Phase 1: Learning

Participants were asked to play the role of medical students as their task was to learn to make a medical diagnosis. They were told that they had enrolled in a course called

Paralymphnal Diagnosis 1A3 and that as part of the course they must learn how to distinguish individuals with a disease called puneria from normal individuals who do not have the disease. To enable them to make this diagnosis, participants were presented with patient profiles, 40 puneria patients and 40 normal patients, in random order. Each profile informed them of whether or not the patient revealed each of 4 symptoms: Dizziness, coughing, fever, and rash. An example is presented in Table 6. They were told that "no individual symptom is perfectly predictive of the diagnosis. That is, no symptom is present in every patient who has the disease and no symptom is absent in every normal patient." The presence or absence of each symptom in puneria and normal patients was determined randomly according to the probabilities listed in the top of Table 7.

Insert Table 6 About Here

Insert Table 7 About Here

Each patient profile was presented to the subjects for 15 seconds after which they were asked to state, by recording on the paper provided, whether or not the patient shown had puneria. They were then given feedback in the form of "No, this patient does not have puneria" or "Yes, this patient has puneria."

After the presentation of these 80 cases participants were told that there was a second disease called zymosis that they should also know how to diagnose. 80 more patient profiles were presented, 40 zymosis patients and 40 normal patients, in random order and participants were again given feedback after recording their diagnosis. Before beginning this second phase of the experiment participants were informed that they should

not forget about the diagnostic scheme they had developed for puneria, but that the diagnostic scheme for puneria was irrelevant to the diagnosis of zymosis. The patient profiles consisted of the same four symptoms, but the probability that any given symptom was associated with zymosis patients or normal patients differed as shown in the bottom of Table 7. Presentation of these 80 cases completed the learning phase of the experiment.

Phase 2: Review

A review phase followed the learning phase during which participants were presented with 80 more patient profiles. The first 40 were meant to remind them of the diagnostic scheme for puneria and consisted of 20 puneria patients and 20 normal patients. The last 40 cases consisted of 20 zymosis patients and 20 normal patients. Participants who failed to achieve 70% diagnostic accuracy during the review phase of the experiment completed the study, but their data were dropped before analyses were performed because I am primarily interested in generalizing this study to judgments that are made by individuals with sufficient expertise to make a reasonable decision. A criterion of 70% was chosen as it is approximately 1 standard deviation below the average diagnostic accuracy achieved (Mean = 77%, St. Dev. = 9%, Range = 48-100%). Twenty-six poor diagnosticians were dropped from the study, leaving 86 participants whose test phase responses were analyzed.

Phase 3: Test Phase

After completing the review phase of this experiment participants were congratulated on completing medical school and were asked to imagine further that they had become specialists who deal exclusively with patients already known to have a paralymphnal illness, of which there are only two kinds: Puneria or zymosis. They were told that "Therefore, every patient you see has either PUNERIA or ZYMOSIS, but not both. That is, your patients have one - and only one - of these illnesses." They were then given the following test of understanding of the instructions, as developed by McKenzie (1998):

> When you see patients from now on, which of the following is true? (a) Each patient has puneria, or zymosis, or neither. That is, patients have either one of the illnesses, or neither of them. (b) Each patient has puneria, or zymosis, or both. That is, patients have either one of the illnesses or both of them. (c) Each patient has puneria or zymosis, but not both. That is, patients have one - and only one - of the illnesses. (d) Each patient has puneria, or zymosis, or both, or neither. That is, patients may have one of the illnesses, both of the illnesses, or neither of the illnesses.

If the participants did not immediately respond (c) then the mutual exclusivity and exhaustiveness of the diagnoses was re-explained and the question re-asked until all participants appeared to understand. They were then told that it was time for them to take their specialists' examination. The test consisted of a series of sixteen patient profiles that were created using every possible combination of the presence/absence of all four symptoms. Participants were told that each patient profile would be presented for 15 seconds during which they should try to form an impression of the patient. After 15 seconds the participants were asked about the patient in one of six ways.

Puneria was the focal diagnosis for half of the participants and zymosis was the focal diagnosis for the other half. That is, half of the participants were asked to rate their confidence that each test patient had puneria and the other half were asked to rate their confidence that each test patient had zymosis. If both groups take the alternative (non-focal) diagnosis into account when making their decision, then the sum of the confidence ratings assigned by the two groups should equal 100. Within each group subjects had the question about their confidence presented to them in one of three ways. Two of the groups were defined by having the question phrased in either an asymmetric or a symmetric manner as in McKenzie (1998). The Asymmetric group was asked "How confident are you that this patient has puneria?" The Symmetric group, in contrast, was asked to rate their confidence in a way that would focus their attention on the non-focal diagnosis: "How confident are you that this patient has puneria rather than zymosis?" To determine whether or not directing attention towards the evidence available to participants would further reduce the

bias observed, a third group was added to this experiment. Participants in the Evidence group were asked to state (by circling the symptoms) which symptoms, if any, favoured puneria and which symptoms, if any, favoured zymosis. Upon completing this task they were asked to assign confidence ratings in response to the Symmetric question.

Results

The average confidence rating assigned to the focal diagnosis was calculated for each of the 16 test cases separately for all six conditions. For each of the Asymmetric, Symmetric, and Evidence conditions the sum of the average probability assigned when puneria was the focal diagnosis and the average probability assigned when zymosis was the focal diagnosis was then calculated for each case. If individuals considered the non-focal diagnoses adequately when assigning their probability ratings, then these numbers should all equal 100. For example, a case with symptoms that are highly indicative of puneria should result in a high rating being assigned to puneria (e.g., 90%) and a low rating being assigned to zymosis (e.g., 10%). A case with symptoms that are highly indicative of both diagnoses (or not at all indicative of both diagnoses) should cause both disorders to be assigned a probability rating of 50% by virtue of there being only two possible diagnoses. This was not observed. Nonetheless, the issue of interest is whether or not manipulating the way in which the question was asked would influence the degree to which the judgments were non-additive (i.e., differed from 100). Cases that are highly indicative of both diagnoses tend to show superadditivity (i.e., P(Puneria) + P(Zymosis) > 100) and cases that are not indicative of either diagnosis tend to show subadditivity (i.e., P(Puneria) + P(Zymosis) < 100) (McKenzie, 1998). So, to avoid superadditive cases and subadditive cases canceling one another out, the degree of non-additivity was determined for each of the 16 test cases by calculating the absolute value of non-additivity. That is, I calculated the

absolute value of the difference between 100 and the sum of the probability ratings assigned to both puneria and zymosis. These calculations resulted in the generation of a single number for each case within each condition that represents the degree to which the participants' judgments were non-additive. The degree of non-additivity averaged across all 16 cases is illustrated in Table 8.

Insert Table 8 About Here

One way ANOVA revealed that the degree of non-additivity was different across these three conditions ($F(2,45) = 9.80$, $MSe = 270.330$, $p < 0.001$). Planned comparison t-tests indicated that the Asymmetric group's judgments deviated from 100 significantly more than did those of the Symmetric group ($t(15) = 2.58$, $p < 0.05$). In addition, the Symmetric group's judgments revealed a significantly greater deviation from 100 than did the Evidence group's judgments ($t(15) = 2.51$, $p < 0.05$). The Evidence group's judgments were not significantly greater than 0 (i.e., they showed non-additivity; $t(15) = 1.40$, $p > 0.10$).

Discussion

These results suggest that the degree to which one under-weights the evidence relevant to a non-focal diagnosis can be reduced by re-evaluating the evidence that is available before making a decision. That is, the degree of non-additivity exhibited by the participants' judgments was reduced if they assigned their confidence ratings after evaluating which symptoms favour each of the diseases under consideration. It is important to note that this extra task did not provide participants with any new information that the other two groups did not have available. It merely directed the Evidence group's attention

towards the diagnosticity of each symptom – an examination that could be performed in all conditions had participants chosen to do so.

One particularly interesting aspect of these results is that the stimuli that were presented were not perceptually ambiguous. There could be no question as to whether or not a particular patient had a rash as the data were presented in the form of simple 'yes' or 'no' responses. The only ambiguity that existed was in the appreciation of the significance of the presence/absence of a particular symptom for each diagnostic alternative: Ambiguity of valence. Combined with the fact that subjects had to evaluate only 4 symptoms and 2 diagnoses, this modest amount of ambiguity should make it less likely that individuals would be biased away from considering non-focal alternatives relative to medical diagnosis in a real-world context. Rarely is the ambiguity inherent in a real-world case so minimal. Physicians may be called upon to comment on a particular diagnosis based on a feature list provided by a chart or description that had been generated by someone else, but more often clinical information must be collected and analyzed in more equivocal settings. The collection of clinical data is likely to be biased by one's preconceptions (Green and Yates, 1995), and the particular features that are identified can be construed differently depending on the diagnosis that one has in mind (Brooks et al., 2000). All of these factors suggest that the effect elicited in this study may actually underestimate the strength of the effect in a more natural context.

At the same time, there are two reasons to be cautious when estimating the generalizability of the current results to actual medical practice. First, the context that was created differed from normal medical practice in that (a) there was a highly controlled diagnostic scheme given that the probability of any symptom was tightly constrained for both diseases, (b) the base rates of occurrence of the two diagnoses were identical, (c) the number of diagnoses that needed to be learned was small, as was the number of symptoms relevant to each diagnosis, (d) the features relevant to each diagnosis were identical and (e)

the number of cases observed within each diagnostic category was large. Any one of these factors might alter the context enough to change the effects observed. Second, the manipulation utilized in the Evidence condition was a strong one that is unlikely to be of use in a real medical setting. In Experiment 4, I instructed participants to focus on the evidence by listing for them all of the symptoms to which they should attend. That is, focus of attention was drawn towards specific features of the case. In most medical situations no one will be present to instruct students to question whether or not a particular symptom supports the differential diagnosis as well as the primary diagnosis. As a result, less directive instruction that could be implemented by the students themselves is likely more appropriate to simulate a clinical setting. In Experiment 5 the issue of evidence consideration was addressed using real clinical cases and actual medical students.

Experiment 5

The issue examined in this experiment was whether or not instructing medical students to re-consider the evidence available to them when making their judgment might better allow them to make unbiased probability estimates. That is, probability estimates that are not influenced to the same extent as the judgments observed in Experiments 1 to 3 by which diagnostic alternative resides in the focus of attention. Rather than having individuals state which particular symptoms favour two specific diagnoses, as was done in Experiment 4, participants in Experiment 5 were simply asked to list for themselves "any evidence that is consistent with diagnosis X." If differential processing is the cause of the unpacking effect, and if instruction to re-evaluate the evidence allows one to avoid differential processing, then I should find that the unpacking effect can be reduced or eliminated through instruction to reconsider the evidence when making a judgment. That is, instruction

to re-consider evidence should lead to more thorough evaluation of implicit non-focal alternative diagnoses.

To test this hypothesis, I conducted a standard unpacking experiment in which participants were shown a case history and were asked to evaluate the probability that the case was representative of either a single diagnosis or two diagnoses. The nature of the instructions to re-evaluate the available evidence was also manipulated. One group was not asked to list the evidence that they detected, but rather was asked simply to assign their probability ratings immediately after working through the case history (Immediate Condition). The other two groups were asked to list the evidence for either the focal diagnosis (Argue for F) or for both the focal and the non-focal diagnoses (Argue for F & NF). It was hypothesized that focusing on the evidence would better enable diagnosticians to take the alternative non-focal diagnosis into account when formulating probability ratings. As such, the prediction was that the unpacking effect would be observed in the Immediate condition, but eliminated in the "Argue for F & NF" condition. The "Argue for F" condition was included to test whether it would produce an exaggerated form of the processing that I believed would predominate in the Immediate condition. If so, then the magnitude of the unpacking effect should be greater in the "Argue for F" condition than in the Immediate condition.

Method

Participants

Nineteen medical students from McMaster University (12 second year and 7 clinical clerks) participated in this experiment. Second year students were contacted through their tutorial leaders, and clinical clerks were recruited to participate at a group session on interviewing for their medical residencies. Those who agreed to participate were handed a

questionnaire booklet and were asked to complete the booklet independent of colleagues and without the use of any reference materials. Return of the questionnaires was arranged by phone contact initiated by the participant, who then received $20 along with feedback regarding both the clinical cases used and the purpose of the study. Nineteen of the 39 (49%) booklets that were distributed were returned.

Materials

The questionnaire booklets consisted of 12 short case histories and a series of questions. After each case history, students were given the names of three diagnoses that could be considered potentially relevant to the case. One of the three diagnoses was randomly selected to be the focal diagnosis while another was used as the non-focal alternative. The third diagnosis was included simply as a distractor. In the packed condition, participants were asked to assign a probability rating to the focal diagnosis and a residual "all other diagnoses" category. In the unpacked condition, participants were asked to assign a probability rating to both the focal and the non-focal diagnoses as well as a residual category. For any given case, one third of the participants read the case history and then were asked to assign their probability ratings (the Immediate condition). The other two thirds of the participants had their focus of attention manipulated in one of two ways. Half of them were asked to list the evidence within the case that was consistent with the focal diagnosis before assigning their probability ratings (the Argue for F condition). The other half were asked to list the evidence that was consistent with the focal diagnosis as well as the evidence that was consistent with the non-focal alternative diagnosis before assigning their probability ratings (the Argue for F & NF condition). The order in which the evidence was requested in the "Argue for F & NF" condition was counter-balanced.

Procedure

In working through all 12 case histories, participants were asked to list the evidence for the focal diagnosis for 4 cases (Argue for F), to list the evidence for both the focal and

non-focal diagnoses for 4 cases (Argue for F & NF), and simply to assign their probability ratings for 4 cases (Immediate). Two of each set of four cases were presented in the packed version, in which a probability rating was required for the focal diagnosis and the residual "all other diagnoses" category. The other two cases in each set were presented in the unpacked version where probability ratings were required for the focal diagnosis, a non-focal alternative diagnosis, and the residual "all other diagnoses" category. After the presentation of each case, all participants were shown a list of three diagnoses and were told that these diagnoses were potentially relevant to the case to ensure that any differences observed between the "Argue for F & NF" condition and the Immediate condition did not arise simply as a result of the experimental materials raising the non-focal alternative (NF).

Results

It was discovered during the course of the study that one of the cases used was inappropriate for many of the second year students as they had not yet covered nephrology. As a result, the analyses reported examine only the remaining eleven cases. Table 9 illustrates that an unpacking effect was observed only when participants were not given an instruction to focus on the evidence (the Immediate condition). Three planned-comparison F-tests were performed, each of which tested a 2 (Unpacking) x 19 (Subject) ANOVA for each level of the instruction variable. Subjects in the Immediate condition judged the focal diagnosis to be significantly more likely when it was presented in isolation relative to when it was presented with the non-focal alternative, $F(1,38) = 4.29$, $MSe = 456.158$, $p < 0.05$. Supportive of my hypothesis, probability ratings did not reveal an unpacking effect when participants were asked to explicitly list the evidence consistent with the focal diagnosis as well as the evidence consistent with the non-focal alternative (the "Argue for N & NF" condition), $F(1,38) = 0.35$, $MSe = 315.803$, $p > 0.5$. Surprisingly, however, the unpacking

effect also disappeared when participants were asked to simply list the evidence in favour of the focal diagnosis (the "Argue for F" condition), $F(1,38) = 0.34$, $MSe = 627.487$, $p > 0.5$.

Insert Table 9 About Here

Discussion

Although the responses in the Immediate condition showed a standard unpacking effect, participants were able to avoid under-weighting non-focal diagnoses when prompted to list the evidence consistent with both the focal and the non-focal alternative diagnoses. In other words, the amount of bias was reduced when attention was directed towards re-evaluation of the evidence before a judgment was required. This was true both when the instructions required subjects to list the evidence consistent with both the focal and non-focal diagnoses and when participants were requested to simply list the evidence consistent with the focal diagnosis. This latter result was unexpected.

The hypothesis I was trying to test is that the unpacking effect arises due to biased processing of the evidence available. That is, when asked to rate just the probability of a single focal diagnosis, judges might turn their attention preferentially to the evidence consistent with that diagnosis and fail to notice (or at least under-weight) available evidence that supports alternative diagnoses. Therefore, I predicted that explicit instruction to list evidence that is consistent with only the focal hypothesis would reinforce the processing that was performed in the immediate condition, perhaps diverting participants' attention even more strongly to just the evidence in favour of the focal diagnosis. The absence of an unpacking effect suggests that this did not occur. However, post-experiment interviews revealed a plausible reason for this unexpected finding.

Many of the subjects in this study viewed the experiment as an educational exercise and requested the chance to work through the cases with their tutorial groups after the experiment was completed. These tutorial sessions provided an opportunity to request information from participants regarding the phenomenology associated with completing the experiment. First, in trying to justify their responses, participants argued that when they were not asked to list the evidence for any diagnoses (the Immediate condition), they felt as though they were heavily influenced by one or two particularly salient features within the case. Using my terminology, they felt as though they were differentially processing the evidence available to them. This subjective report supports my hypothesis that differential processing causes the unpacking effect in that this condition was the only one that exhibited an unpacking effect. Second, with respect to the unexpected lack of an unpacking effect in the "Argue for F" condition, numerous participants claimed that it was difficult for them to answer the question "please list any evidence that is consistent with diagnosis F" without also generating evidence that was inconsistent with the focal alternative. That is, asking subjects to list the evidence in favour of the focal diagnosis appears to have made it more likely that participants would also more fully consider the evidence consistent with the non-focal alternative diagnosis. So, subjects in the "Argue for F" condition appeared to have spontaneously aligned themselves with those in the "Argue for F & NF" condition. Although this was not the intended effect of these instructions, it does account for the elimination of the unpacking effect in the "Argue for F" group. Furthermore, this result strongly suggests that instruction to re-evaluate the evidence at the time of judgment need not include the explicit mention of a diagnostic alternative in order to get students to consider non-focal diagnoses. Instructions that do not rely on the explicit mention of a particular diagnosis are likely to be most useful for students during everyday practice.

So, it appears that instruction to re-evaluate the evidence at the time of decision can eliminate bias created by the explicit presentation of a diagnostic suggestion. To this point,

however, directing subjects to re-evaluate the evidence at the point of decision has been confounded with instructions to generate the evidence that is relevant to the diagnoses. In other words, I can not say for certain that the critical factor in elimination of the unpacking effect is re-evaluation of the evidence at the point of making a judgment rather than simply undertaking a more deliberate approach to evidence collection. Experiment 6 represents an attempt to tease apart these two possibilities.

Experiment 6

Imagine a consumer who is faced with a difficult decision such as which refrigerator to purchase. The decision is difficult because there are many aspects to consider, including size, colour, energy efficiency, shelving arrangement, and price. When faced with such a problem, most people will collect information on many different models and try to compare each one against the others. Come the end of the search, however, when the consumer refocuses his or her attention on a particular model and has to make a final decision, there is no guarantee that he or she will use all of the information that was so carefully collected. I have argued up to this point that the major process that allows the unpacking effect to occur is focusing attention on the evidence consistent with a focal hypothesis at the moment of action. Experiments 3 and 5 revealed that it is possible for judges to have several diagnoses available and still make judgments that reveal the unpacking effect if the judges do not focus their attention on the evidence. However, if it is true that the major process responsible for the unpacking effect is focusing attention on one hypothesis at the moment of action, then it should also be possible to show that simply having the evidence available (i.e., having undertaken the evidence generation process) will still allow the unpacking effect unless the evidence is re-evaluated at the point in time at which the decision is to be made.

To test this hypothesis, I asked undergraduate psychology students to attempt to solve a series of mystery stories. Such a task can be viewed as analogous to a medical diagnostic decision-making task as the decision maker must select particular relevant features from a mass of information and then use these features to weigh the probability of multiple alternatives. As participants read each story, they were asked to try to determine "whodunit." That is, they were to solve the crime that took place in the story. After reading the story to completion, all participants were asked to perform an evidence generation task that required them to write down everything they could think of that might convince someone else that (a) the focal character committed the crime and (b) a non-focal alternative character committed the crime. I then manipulated the delay between the evidence generation task and the assignment of probability judgments. As all participants engaged in the evidence generation task, any differences across groups in the magnitude of the unpacking effect can not be attributed to one group having taken more care to evaluate the evidence than another group. All that differed was the point in time at which the evaluation took place relative to when the probability judgment was required.

Method

Participants

Sixty students enrolled in a first year Introductory Psychology course at McMaster University participated in this experiment for course credit. Subjects were run in groups of 5 to 10 students.

Method

Each subject read a series of 6 mystery stories with the task of trying to determine who committed the crime. Order of presentation of the stories was counter-balanced. Three stories were one page in length while the other three were slightly longer at two pages. Two

characters were randomly chosen from the story and one of these two was designated the focal character. As in past experiments, half of the subjects for each story were asked to assign a confidence rating indicating how likely they thought it was that the character named (the focal character) committed the crime. The other half of the participants were asked to assign similar confidence ratings to both the focal character and a non-focal alternative character. The presence of an "all other characters" category was used to ensure that the probability judgments assigned summed to 100.

To ensure that each participant thought about the evidence available in the story, they were instructed, after reading each story, to write down anything that they could think of that might convince someone else that the focal character committed the crime. They were then asked to do the same thing for a second non-focal alternative character. To determine whether focusing attention on the evidence at the point at which a judgment is required is critical to eliminating judgment bias, the delay between consideration of the evidence and the assignment of probability ratings was manipulated. After arguing in favour of both the focal and non-focal characters for the first and second stories read, participants simply moved on to the next story. After doing the same for the 3rd and 6th stories participants were given a memory test in an attempt to re-focus their attention on the focal character. Subjects were asked to recall everything they could about the focal character, including any information that they wished they knew, but had not been told. They were then asked to assign their confidence ratings (the Distracting condition). In contrast, after reading the 4th and 5th stories, participants were asked to assign their confidence ratings immediately after arguing in favour of both characters (the Immediate condition). Finally, after completion of the 6th story participants were asked to recall the first and second stories that they had read as well as possible. For both stories, subjects were then asked to assign their confidence ratings for either the focal character alone or for both the focal and non-focal characters (the Delay condition).

The time that each participant took in each phase of the experiment was controlled. Participants were told that they could not move on to the next phase of the experiment until the experimenter told them to do so. This was done to ensure that each participant put some effort into the evidence generation task. Participants were given 3 or 5 minutes to read each story (depending on whether the story was 1 or 2 pages long). They were then given 1.5 or 2 minutes to perform each of the other tasks – arguing in favour of particular characters and responding to the distracting memory task. They were not limited in the time it took to assign their probability ratings.

Results

A series of planned comparison ANOVAs supports the notion that engagement in evidence generation is insufficient for reducing the response bias illustrated by the unpacking effect. Rather, the evidence must be considered at the point at which the decision is made. As illustrated in Table 10, the unpacking effect was eliminated when participants assigned their confidence ratings immediately after formulating their arguments for the two characters (Immediate condition), $F(1,118) = 1.2$, $MSe = 637.626$, $p > 0.25$. This result replicates that of the "Argue for F & NF" condition of Experiment 5. However, when a delay was introduced between evidence generation and the assignment of confidence ratings, the unpacking effect returned, $F(1,118) = 5.7$, $MSe = 633.055$, $p < 0.02$. Re-focusing participants' attention on the focal character after they listed the evidence by imposing a memory test resulted in an unpacking effect that was intermediate in size relative to the other two conditions, but non-significant, $F(1,118) = 2.4$, $MSe = 551.880$, $p > 0.1$.

Insert Table 10 About Here

To further tease apart the effect of the delay conditions, one-way ANOVAs were also performed independently for each of the levels of unpacking. When participants were asked to assign confidence ratings to both the focal and the non-focal diagnoses (the unpacked condition) the delay manipulation had no effect. That is, there was no difference in the probabilities participants assigned the focal diagnosis regardless of whether the probabilities were assigned immediately after evaluating the evidence or after a delay. $F(2,177) = 2.2$, $MSe = 557.866$, $p > 0.1$. A different pattern was observed when the non-focal diagnosis remained implicit (the packed condition). When participants were asked to assign a rating to only the focal diagnosis, the rating assigned was lower in the Immediate condition than it was in the Delay condition, $F(2,177) = 4.7$, $MSe = 657.175$, $p < 0.02$. Post-hoc Tukey tests revealed that responses in the distracting condition did not differ significantly from either of the other groups when only the focal diagnosis was rated.

Discussion

The absence of an unpacking effect in the group that assigned confidence ratings immediately after evaluating the evidence again supports my hypothesis that the unpacking effect is often driven by differential processing of the evidence available to a diagnostician. That is, directing attention towards the evidence can reduce the tendency to under-weight implicit diagnoses. The fact that the unpacking effect occurred when there was a delay between the evaluation of the evidence and the assignment of confidence ratings suggests, however, that it is critical that attention be focused on the evidence at the point in time when a decision is required. Undertaking controlled evidence collection was insufficient to allow participants to avoid the bias created by the explicit presentation of a diagnostic suggestion unless that evidence was evaluated at the time of judgment.

A memory test was used in the Distracting condition in an attempt to re-focus participants' attention on the focal diagnosis immediately after the evidence generation task. I hypothesized that one's attention can be drawn away from the evidence he or she has collected relatively quickly, thereby re-creating the context within which under-weighting is typically observed. While an intermediate sized unpacking effect was observed in this condition, the difference in the probability assigned to the focal diagnosis in the packed and unpacked versions of the study did not reach significance. The hypothesis remains plausible, however, even though the data collected does not support it conclusively. A memory task such as this one is likely a weaker re-orienting task than the ones that exist in medicine. Medical diagnosticians might be distracted by a different case or they might be asked to comment on whether or not they would like to see a particular diagnostic test performed. Both of these tasks are likely to elicit more cognitive effort than an artificial memory task. As a result, both have the potential to weave themselves into the more natural flow of events, thereby making it less likely diagnosticians will return to the mental state that was present before the re-orienting task arose.

In summary, the experiments presented in this chapter support the idea that the bias produced by unpacking is largely the result of differentially focusing one's attention on the evidence at the moment of making an explicit judgment. Up until now it has been assumed that processing evidence differentially entails focusing attention on one or two features that are particularly relevant to the diagnosis that a judge is asked to evaluate. Chapter 5 describes a study that was conducted to more directly examine the influence of over-weighting features on diagnostic judgments. Specifically, I attempted to manipulate the mnemonic effectiveness of the features by altering the terminology used. The goal was to determine whether features that are mnemonically effective will be weighted more heavily than less salient features when assessing diagnostic possibilities.

CHAPTER 5

Over-weighting mnemonically effective features

Experiment 7

When clinical features are presented to a diagnostician using medicalese (i.e., technical descriptions that utilize medical terminology), the informational content they provide might be weighted more heavily in the decision-making process relative to when the same features are presented using lay terminology. That is, the use of medicalese might increase the medical familiarity of a feature, thereby resulting in it being given more weight when a diagnostic decision is required.

This is not necessarily the case, because the information contained in a medical term (e.g., dyspnea) is equivalent to that provided by a lay synonym (e.g., shortness of breath). However, the encoding specificity principle, identified by Tulving (1972) while studying human memory, provides one explanation for why the use of medical terminology might produce such an effect. This principle states that items in memory will be better recalled when the context within which memory retrieval takes place matches the context within which learning occurred (see also Morris, Bransford, and Franks, 1977). In discussing cases with colleagues or students, physicians regularly translate a patient's complaints from lay terminology into medicalese (i.e., technical descriptions that utilize medical terminology). As a result, a new case will better match the context established during learning when medical language is used relative to when lay terminology is presented. In turn, features that are presented in medical terminology should act as more effective cues to trigger particular diagnoses and should be more memorable after a case has been evaluated.

I hypothesize that similar factors cause the differential processing that occurs when participants evaluate cases within the context of a particular diagnostic hypothesis. The presence of a diagnostic suggestion might cue an individual to attend to particular features of a case, thereby making them more salient. This could, in turn, cause such features to be over-weighted during the decision-making process relative to features that are consistent with alternative hypotheses.

One theory of expertise in medicine has proposed that students who utilize a form of medicalese called semantic qualifiers (e.g., acute, distal, sudden) are more likely to accurately diagnose clinical cases (Bordage and Lemieux, 1991). Although substantial data have been presented in support of this correlation, it is not clear whether accuracy improves as a result of the use of these qualifiers or whether the better students are simply more able to describe the cases using medical nomenclature. If it is true that the use of semantic qualifiers can improve diagnostic accuracy, then manipulating the way in which the case history is presented should also influence the diagnostic conclusions that are reached. More specifically, the more medical nomenclature is used to describe the features consistent with a particular diagnosis, the more likely it should be that a diagnostician will conclude in favour of that diagnosis.

Method

Participants

Eleven first and second year family medicine residents from the University of Toronto and three second year residents from McMaster University participated in this study during two group sessions. Informed consent was obtained and the groups were given $30 for each resident that participated to be used at their discretion. At the end of the

experimental session all participants were provided with feedback regarding both the purpose of the study and the clinical cases used.

Materials

Six clinical cases were selected that maintained the following properties. First, these cases were developed so that two diagnoses could be viewed as equiprobable and plausible (Cunnington et al., 1997). Second, for each of the two diagnoses, at least three features were present within the case that could be manipulated for the present purpose. That is, at least three features could be presented using either medical or lay terminology. This was manipulated in one of four ways. (1) A lay description could be converted to medicalese (e.g., shortness of breath = dyspnea). (2) A semantic qualifier could be absent or present (e.g., chest pain vs. retrosternal chest pain). (3) An interpretation of clinical data could be provided (e.g., WBC = 12,000 vs. WBC elevated (12,000)). (4) The absence of particular symptoms could be left implicit or mentioned explicitly (e.g., breathing was normal = there has been no wheezing or hoarseness). Note that only the second of these four manipulations was argued by Bordage and Lemieux to provide an indication of expertise. Four manipulations were used, however, to increase the flexibility with which the features could be made more or less mnemonically effective. A full list of the features that were manipulated is presented in Table 11. Note that the features listed do not represent a comprehensive list of either the entire case history or all of the features that were intended to be indicative of diagnosis A or diagnosis B. They are simply the features that were manipulated across version of the case history.

Insert Table 11 About Here

Two versions of each case were created by manipulating the features that were consistent with Diagnosis A or Diagnosis B. In one version, the features consistent with

Diagnosis A were presented using lay terminology while the features consistent with Diagnosis B were presented using medicalese. In the other version the reverse was true. The features consistent with Diagnosis A were presented using medicalese while the features consistent with Diagnosis B were presented using lay terminology. For each case, each version was presented to half of the participants.

Procedure

Participants were free to work through the test booklet at their own pace. They were told, however, that they should not turn each page until they were committed to proceeding as they were not allowed to turn back to earlier pages. This instruction was given to ensure that (a) participants read the case histories carefully before moving on to the questions specific to each case, and (b) the number of tasks undertaken between reading the case and the memory test for that case was consistent across condition. After reading the first case history, participants turned the page and were asked "Based on the information that you have just read, please list the two diagnoses that you feel to be most likely." On the next page they were shown Diagnosis A and Diagnosis B and asked to rate the probability that the case they had just read was representative of these diagnoses. Participants were told that the patient had only one disorder and that, therefore, the inclusion of an "all other diagnoses" category required that the responses provided sum to 100. They were then asked to rate their confidence in their probability ratings using a 100-point scale. This procedure was repeated for the remaining 5 cases.

After the 6th case, participants were given a memory test. They were told, for example, that "the first case that you were shown focused on a 48 year old woman with epigastric abdominal discomfort." Participants were asked to recall everything that they could about the case. Afterwards a cued recall test was performed in which participants were reminded of the two diagnoses for which they had assigned a probability and were

asked if these diagnoses brought to mind any additional information. These two memory tests were then repeated for the remaining 5 cases.

Results

Hypothesis Generation

Participants in the medicalese condition were as likely as those in the lay terminology condition to generate each of the diagnoses associated with each case. The diagnosis whose features were presented using medical terminology was named 45 out of a possible 84 times (14 participants x 6 cases) whereas the diagnosis whose features were presented using lay terminology was named 42 out of 84 times. One could argue that the first diagnosis named is the hypothesis that is predominant in the participants' mind, but there was also no difference across condition of which diagnosis was named first (25 vs. 26 cases for the medicalese versus lay descriptions, respectively). These results are not terribly surprising as the cases were developed to be equally indicative of both diagnoses. So, to some extent this analysis serves as a manipulation check as it proves that the information presented in the medicalese condition was equivalent to that presented in the lay terminology condition. Hence, any differences observed across condition in the probability judgments are not due to there simply being more information presented in the medical version relative to the lay version of each diagnosis.

Probability judgments

The participants' belief in a diagnosis, as expressed in the probability ratings, was greater when the features that were indicative of that diagnosis were presented using medical terminology relative to when those same features were presented using lay terminology. Diagnoses whose features were presented using medical terminology were assigned an average probability rating of 46.5 (standard deviation = 24.532). When the same

information was presented using lay terminology the probability rating assigned to the same diagnoses dropped to 35.5 (standard deviation = 23.761), $t(83) = 2.31$, $p < 0.05$.

Twelve separate comparisons can be made to ensure the robustness of this result by examining the probability ratings assigned to each diagnosis across version (6 cases x 2 diagnoses). Table 12 illustrates the mean probability assigned to both diagnosis A and diagnosis B as a function of whether each diagnosis was presented using a technical or lay description of their features, for each case independently. Within each case, the middle column indicates the probability assigned to diagnosis A. Table 12 reveals that subjects believed the probability of these 6 diagnoses to be greater when the features consistent with diagnosis A were presented using the technical description relative to when they were presented using the lay description. This was true for every case except case 3 (panel c). The same can be said for diagnosis B. The last column of each sub-table indicates for each case that diagnosis B received a greater probability rating when the features consistent with diagnosis B were presented using a technical description relative to when the same features were presented using a lay description. Again, case 3 (panel c), was the lone exception. So, 10 out of 12 comparisons revealed the predicted pattern of results ($p < 0.02$ using a Sign Test).

Insert Table 12 About Here

Memory Test

Finally, Table 13 illustrates that the language used to describe a clinical case will influence the memorability of the features in that case. The first column of data shows that the features that were manipulated across version of the case history (i.e., those listed in Table 11) were more likely to be recalled when they were presented using medicalese

relative to when they were presented using lay terminology, $\chi^2 = 18.9$, $p < 0.01$. This was not the only effect that medicalese had on the memorability of the cases. As mentioned earlier, not all of the features that were consistent with one diagnosis or the other were manipulated across the two versions of the case history. The second column of data in Table 13 illustrates that presenting features that are consistent with a particular diagnosis in medicalese increases the memorability of all of the features that are consistent with that disease. That is, exactly the same features, presented the same way in both versions of the questionnaire were recalled more often when the other features consistent with the same diagnosis were presented using medicalese relative to when they were presented using lay terminology, $\chi^2 = 4.7$, $p < 0.05$.

Insert Table 13 About Here

Discussion

Lingard and Haber have argued that "[t]he essence of physician-to-physician communication is deciding what is worth saying – and what is not" (1999, p. S124). The current study extends this statement in that whether or not something is perceived as worth having been said might be influenced by the way in which it was said. A resident who is told of a particular symptom using medical terminology appears to be more likely to view that feature as being clinically "relevant" relative to when the same symptom is presented using lay terminology. This was evidenced by the finding that presenting clinical features by using medicalese rather than lay terminology increased the probability that residents assigned to the relevant diagnosis. Furthermore, the use of medicalese increased the likelihood that residents remembered those features. Presumably the effect on memorability

is a precursor to the influence on probability ratings, but I can not make that claim conclusively based on the design of this study.

This finding supports Bordage and Lemieux's claim that the use of semantic qualifiers can improve diagnostic ability (1991). More precise statements of the clinical features may serve as cues to particular diagnoses. Medical personnel who detect and take advantage of such cues should, therefore, be more likely to arrive at the correct diagnosis.

In the context of this dissertation, this study supports the hypothesis that variables that influence the mnemonic effectiveness of features within a case will affect the judgment process by altering the perceived probability of potential diagnoses. This was the case even though the informational value of the features presented in the two conditions was identical; that is, the two diagnoses were equally likely to be named as potential diagnoses. By extension, it seems plausible that judging an explicitly presented diagnostic suggestion might bias one's memory or construal of a case. This memory bias could cause the evidence available to be processed differentially, leading to under-weighting of the evidence consistent with alternative non-focal diagnoses. Whether non-focal diagnoses continue to be under-weighted in this way when diagnosticians themselves generate the focal diagnosis will be examined in Chapter 6.

CHAPTER 6

Self-generated focal diagnoses

Experiment 8

As mentioned in the introduction, many medical cases are evaluated in the context of a diagnostic hypothesis. Clinicians or students are often presented a case with instructions to comment on the likelihood that it is representative of a particular diagnosis. In this chapter, I address whether diagnosticians are susceptible to the same under-weighting of non-focal alternative diagnoses that was observed in Experiments 1 to 6 when they themselves generate the focal diagnosis.

Previous work by Koehler (1994) suggests that participants are more likely to appreciate non-focal alternatives when they themselves are responsible for generating the focal hypotheses. This may occur because participants who generate their own focal diagnoses make their judgments without any cue that causes them to process evidence differentially. That is, attention is not drawn preferentially toward the features consistent with a particular diagnosis, thereby allowing for the full consideration of alternatives. Koehler asked undergraduate psychology students to generate a list of potential Academy Award winners. They then selected the three hypotheses that they viewed as most likely and assigned a probability rating indicating how likely they thought it was that one of these three would prove to be the eventual winner. The same three hypotheses were also provided to yoked subjects who had not undertaken the hypothesis generation phase of the experiment. The yoked subjects reliably rated the three alternatives as more probable than did the participants who generated the hypotheses for themselves. Koehler's explanation was that

subjects who engaged in the hypothesis generation phase of the experiment made their judgments within the context of multiple alternative hypotheses, thereby making them more conservative in the probability ratings they were willing to assign. In other words, subjects who generate the focal hypotheses for themselves are less likely to under-weight the implicit non-focal hypotheses that remain part of the residual.

The outcome of Experiment 6 suggests that this effect should be transient. Koehler's participants who undertook the hypothesis generation phase of the experiment were likely able to avoid under-weighting the non-focal alternative hypotheses as they had no cue that would have led them to process the available evidence differentially. The results of Experiment 6 suggest, however, that it is insufficient to simply undertake a fair evaluation of all of the evidence. Rather, judges must evaluate the evidence in this way while engaged in the final judgment process. As such, participants who undertake an hypothesis generation phase should be as likely to under-weight alternative diagnoses after the passage of time as are those who skip the hypothesis generation phase altogether. Re-presentation of the focal diagnosis that was generated should serve to reinstate a cue that causes differential processing. To test this hypothesis, Experiment 3 was repeated without the explicit presentation of a diagnostic suggestion. Given the preceding argument, it was predicted that participants would view their own hypotheses as more probable when a delay is introduced between the hypothesis generation phase and the probability assessment.

Method

Participants

The participant pool for this study consisted of second year medical students from McMaster University's graduating class of 2002. A sample of tutorial leaders asked their students if they would participate. Those who agreed were run through the experiment in

their tutorial groups during two sessions separated in time by an average of 10 days (range = 7 to 14 days). 25 students participated in 5 groups. Upon completion of the second group session participants were paid $20 and provided feedback regarding both the clinical cases used and the purpose of the study.

Materials

Participants were presented 10 case histories, each of which was followed by a series of 5 questions. (1) Given the case history that you have just read, please assign a primary diagnosis that you believe is the cause of the client's complaints. (2) Please assign a number between 0 and 100 indicating how likely you think it is that the case is representative of the diagnosis that you have provided (0 = no chance, 100 = definitely this diagnosis). (3) Are there any diagnostic tests that you would like to see performed to aid you in your decision? If yes, please list them. (4) While reading the case history did you consider any diagnosis apart from the one that you named above? If yes, please state the diagnosis that you consider to be the most likely differential. (5) Please rate the typicality of this case on a scale of 1 to 100. That is, how typical is this case history of all cases with the diagnosis that you believe to be most plausible.

Procedure

During the first pass, each participant worked through all 10 case histories and assigned responses to all five questions for each case. One or two weeks later each participant was shown the same 10 cases. This time the first and second questions were replaced with "Given the case history that you have just read, please assign a number between 0 and 100 in the space below indicating how likely you think it is that the case is representative of the given diagnoses (0 = no chance, 100 = definitely this diagnosis)." The diagnoses listed were the diagnoses that were generated by the participant in response to questions 1 and 4 during Pass 1. If during pass 1 no differential diagnosis was generated in response to question 4, participants were simply presented with the primary diagnosis

that they generated in response to question 1 and asked to assign a probability rating to this diagnosis alone. Otherwise the questionnaires used during the two sessions were identical.

Results

In completing all ten cases, 246 observations were generated that could be analyzed for the unpacking effect – in 4 of the 250 cases the participant failed to assign a probability rating to the primary diagnosis. Table 14 presents the average probability assigned to the focal diagnosis as a function of both pass and whether or not an alternative diagnosis was generated in response to question 4 during pass 1. A 2 (Session) x 10 (Case) ANOVA was performed. A significant effect of 'session' revealed that the probability assigned to the focal diagnosis was higher during the second session than it was when participants initially generated the diagnosis during pass 1, $F(1, 238) = 15.325$, $MSe = 216.075$, $p < 0.001$. This occurred despite the fact that the non-focal residual category was unpacked during pass 2 relative to pass 1, a manipulation that should operate in a direction opposite to the effect observed. There was no significant effect of 'Case' ($F(9,238) = 0.541$, $MSe = 518.957$, $p > 0.8$).

Insert Table 14 About Here

When no alternative diagnosis was reported in response to question 4 during pass 1, participants were still asked during pass 2 to assign a probability rating to their primary diagnosis. In other words, their task was identical to that performed during pass 1 with the exception of not being explicitly asked to undergo hypothesis generation. As can be seen by comparing the first and second columns of data in Table 14, the probability rating assigned increased during pass 2 regardless of whether or not a differential diagnosis was

generated and presented for evaluation during pass 2. A 2 (Session) x 2 (Alternative Diagnosis: Generated vs. Not Generated) x 10 (Case) ANOVA was performed and did not reveal a 'session' x 'alternative diagnosis' interaction, $F(1,231) = 0.787$, $MSe = 215.975$, $p > 0.4$. Although the low number of observations in the 'not generated' group suggests that there may not be enough power to observe a significant interaction, the consistency of the increase observed from pass 1 to pass 2 regardless of whether an alternative was generated indicates no trend at all toward an interaction.

I also examined whether or not the phenomenon being illustrated by the probability ratings might be shown to influence management strategies by asking the participants to list the diagnostic tests that they would be interested in seeing performed. Indeed, participants requested significantly more tests during pass 2 (mean = 2.97) than during pass 1 (mean = 2.75), $F(1,238) = 7.10$, $MSe = 0.817$, $p < 0.01$.

Finally, the effect of the number of alternatives presented on typicality ratings was analyzed. Again, a significant effect of 'session' was observed as the typicality ratings were greater during pass 2 (mean = 68.08) than during pass 1 (mean = 64.28), $F(1,227) = 6.964$, $MSe = 249.575$, $p < 0.01$.

Discussion

The increase in probability assigned during pass 2 relative to pass 1 suggests that subjects were more likely to take non-focal alternative diagnoses into account when they themselves generated the focal diagnosis in the absence of an explicit diagnostic suggestion. This result was predicted because the absence of a diagnostic suggestion eliminates one cue that might cause individuals to process the available evidence differentially. This non-biased evaluation of the evidence is transient, however, further supporting the idea that

diagnosticians should be instructed to re-evaluate the evidence consistent with multiple hypotheses during the act of decision-making.

Two alternative explanations might account for the increase in the probability assigned to the focal diagnosis during pass 2 relative to pass 1. First, the explicit presentation of diagnostic suggestions during pass 2 may have led participants to mistakenly assume that the hypotheses were chosen for presentation due to their being probable diagnoses, thereby causing subjects to inflate their probability estimates. This 'prestigious suggestion' explanation is unlikely, however, given that every participant interviewed after the experiment claimed to have noticed that the diagnostic hypotheses presented were the ones they had generated the previous week. It is possible that participants recognized the diagnoses as being ones they themselves generated, but used the subsequent diagnostic presentation as confirmation of their initial hypotheses. Phenomenological reports suggest that this was not the case, but I can not make this claim conclusively.

The second alternative explanation is more difficult to rule out given the procedure used. It may be that the hypotheses generated by participants during the first pass struck them as being surprisingly fluent when presented explicitly one week later. The strong sense of "fit" that resulted might have caused participants to increase the probability they assigned to the focal diagnosis. Supportive of this hypothesis is the finding that the probability assigned to the focal diagnosis increased even when an alternative differential diagnosis was not generated. For these cases participants claim not to have been considering any diagnosis other than the focal diagnosis during pass 1. That is, the number of overtly identified diagnoses being evaluated should have been the exact same during both passes, yet the probability assigned changed. This result can easily be reconciled with my hypothesis that participants' judgments are influenced by the degree to which they consider the alternatives with the realization that considering non-focal alternatives during pass 1

might have consisted of evaluating diagnoses that remained implicit. That is, non-analytic processes might play a role.

To this point I have discussed consideration of non-focal alternatives in terms that imply an analytic process involving consideration of the evidence for specific non-focal alternatives. However, it is also possible that judges maintain a sense that there are a number of plausible alternatives even when these alternatives are not overtly identified. In other words, participants who did not name an alternative diagnosis during pass 1 may have been aware that alternatives did exist, even though they could not name those alternatives. If the explicit presentation of a diagnostic suggestion cued individuals to focus their attention on the evidence consistent with that suggestion during pass 2, then an increase in the probability rating assigned would be expected if the non-specific awareness was not reinstated. Chapter 7 presents a study that examines the possibility that such non-analytic processes influence the judgments reached during a subjective probability task. That is, I attempted to determine whether people's judgments are subject to change based on inferences that non-focal alternatives are potentially available, but not readily accessible.

CHAPTER 7

Non-analytic influences on probability ratings

Experiment 9

While most of the work presented up to this point has focused on the role of generating and evaluating specific alternative diagnoses, it is not necessary to generate specific alternatives in order to have a sense of the number of hypotheses that could potentially be considered. This chapter introduces the idea of implied numerosity to suggest that one component of probability evaluation involves an assessment of the size of the category to which an event belongs. That is, the context of a question, including the alternatives offered explicitly, may alter a judge's impression of the number of possible alternatives that have not been mentioned.

Although descriptions of the phenomena described in the preceding chapters might cause humans to appear to be illogical decision-makers, the implied numerosity framework suggests that some part of that appearance might be the by-product of a rational judgment process. That is, the experimental manipulations that have been shown to alter probability ratings may have done so, at least partially, by changing the perceived size of the category being evaluated. This proposal will be argued for in two ways: By an assessment of its ability to explain results already present in the literature and by examining novel predictions that have been tested and are presented here. First, however, I briefly discuss the need for such a proposal.

Much of the work performed in the area of subjective probability focuses on the effect of packing and unpacking specific alternatives. Recall that Tversky and Koehler

(1994) speculated that the amount of support in favour of any one alternative is constrained by both memory limitations (i.e., the ability to recall alternative hypotheses) and attentional capture (i.e., an increased salience as a result of an alternative's explicit presentation). However, this focus on specific hypotheses assumes that only explicitly considered alternatives are influential. Surely, there is a more generic, exemplar-free way of performing frequency estimations that can affect judgments of probability. Just as one can estimate the size of a choir by trying to pick out multiple specific voices, the volume of the chorus as a whole must also provide valuable information. The alternatives that are explicitly presented could possibly provide information that alters a judge's perception of the number of alternatives that could be generated. That is, the alternatives themselves imply numerosity.

This possibility is consistent with personal experience as well as with discussion held with experimental participants. Ask yourself the question "How probable is it that Russia hosts the world's largest prison?" Although capable of generating a long list of countries, many of which would be reasonable alternatives (including China, which is the correct answer), most people do not attempt to do so spontaneously. In fact, the work presented throughout this dissertation suggests that even trained diagnosticians may not generate (or at the very least, they under-appreciate) the most likely non-focal diagnoses unless they are explicitly mentioned. Rather, a more commonly observed pattern is the automatic consideration of the explicitly mentioned alternatives (or at most, 1 or 2 alternatives that are included only implicitly yet come to mind quickly) followed by some vague consideration of the question as a whole. It seems unlikely that this discounting of alternatives not explicitly mentioned is due to low motivation during the experimental situation, because this strategy has been observed in the most dedicated subjects as well as during natural conversation with other individuals. Rather, when asked to evaluate the likelihood that a given answer is correct (or that a given event might occur) there seems to be a natural tendency to make a decision on the basis of some general impression of its

probability rather than through the use of specific comparisons of a number of possible alternatives.

What creates this "general impression" that seems to be driving the responses given by subjects? Given that people do not expend a lot of effort generating additional alternatives to compare with the explicitly given alternatives, what determines the magnitude of the probabilities assigned? I propose that the amount of 'support' one holds in favour of the residual category is determined, in part, by a general impression of the size of the category that judges are asked to evaluate. As the number of potential hypotheses increases, the probability of any one hypothesis being correct decreases. Hence, unpacking effects might arise, at least in part, by the explicit presentation of component hypotheses altering the judge's perception regarding the number of alternatives that exist. If presenting three diagnostic alternatives leads an individual to believe that a greater number of potentially relevant diagnoses exist (even if the diagnoses themselves are inaccessible) relative to when just one diagnostic alternative is presented, then the probability assigned to any one diagnosis (including a focal diagnosis) should decrease.

This notion that implied numerosity is one mechanism whereby 'support' can be generated yields two predictions that will be illustrated by use of an example. First, the size of the category required to include all unpacked alternatives should systematically alter the magnitude of the unpacking effect. Consider again the question "what country hosts the largest prison in the world?" If subjects are sensitive to implied numerosity, then the decrease in the probability assigned to the focal alternative (Russia) should be greater if the residual is unpacked into alternatives which include countries from all over the world (e.g., United States and Australia) relative to when the unpacked alternatives are all Asian countries (e.g., China and Japan). The category size (i.e., the number of possible alternatives) in the latter is smaller. If probability ratings are sensitive to perceived category size, then any individual country within this latter context should receive a larger probability

rating as the probability available to be assigned needs to be spread over fewer implicit alternatives. Therefore, focal hypotheses should be rated as more probable, resulting in a smaller unpacking effect when components of the residual category are stated explicitly.

The second prediction is that if the alternatives that are unpacked are held constant, then there should be a greater unpacking effect the smaller the packed category is perceived to be. Again, the smaller a category is perceived to be, the greater should be the probability assigned to any one alternative, and so the greater the effect should be upon unpacking additional alternatives. Providing a hint that the country which hosts the largest prison in the world is in Asia should serve to reduce the size of the category of focus in the packed condition. Therefore, when a hint is given, Russia should receive a larger probability rating -- and unpacking into China and Japan should result in a greater difference in probability ratings -- relative to when no hint is given.

Method

25 undergraduate Introductory Psychology students were presented with 50 trivia questions and asked to evaluate what percentage of people they thought would generate particular answers in response to being asked each question. Participants were asked to assign a number between 0 and 100 to each alternative and told that the inclusion of an "all other alternatives" option should result in the numbers summing to 100 for each question. They were also told that they should not assume that the correct answer is necessarily one of the hypotheses presented.

Table 15 illustrates, using an example, the five experimental conditions described below. Trivia questions were presented in one of five conditions, all of which had a focal hypothesis consistently presented: (1) a Large Category (i.e., no hint) Packed condition (LP), (2) A Small Category (i.e., hint given) Packed Condition (SP), (3) a Large Category

Unpacked Similar Condition (LUS) in which the alternatives that were unpacked were all from a relatively small category, (4) a Large Category Unpacked Dissimilar Condition (LUD) in which the unpacked alternatives came from a wide range of possible alternatives, and finally, (5) a Small Category Unpacked Condition (SU) in which the alternatives were the same as in the LUS condition and a hint was provided. Only the focal alternative was presented in the Packed conditions while three alternatives, including the focal, were presented in all unpacked cases.

Insert Table 15 About Here

Results

Tables 16 and 17 illustrate the scores assigned to the focal diagnosis in each condition averaged across 49 questions - one question was dropped from analysis as an error was found in the alternatives list presented. Table 16 shows that, as predicted by Support Theory, the probability assigned to the focal diagnosis was higher in the Packed Condition than in either Unpacked conditions. Furthermore, as the implied numerosity framework predicted, the decrease in the probability assigned to the focal was greater when the unpacked non-focal alternatives were Dissimilar (i.e., taken from a large category) relative to when the unpacked non-focal alternatives were Similar (i.e., taken from a small category). That is, the difference between the LUS condition and the LUD condition is statistically significant ($t(48) = 2.25$, $p < 0.05$) using a repeated measures, two-tailed t-test with question as the unit of analysis.

Insert Table 16 About Here

Table 17 reveals that the second prediction made by the implied numerosity framework was also observed to be true. Keeping the unpacked alternatives constant, but varying the size of the category of focus in the Packed condition resulted in a larger unpacking effect when the category size in the Packed condition was small (SP – SU) relative to when the category size of the Packed condition was large (LP – LUS). This difference was significant ($t(48) = 1.98$, $p < 0.05$) using the same analysis as above. It can also be observed in Table 17 that the difference in the two unpacking effects resulted from an increase in the probability assigned to the focal diagnosis in the Small category Packed condition. This was also predicted by the implied numerosity framework, because the similarity of scores in the two Unpacked conditions is to be expected if the manipulation of unpacking into similar alternatives had the same effect as providing a hint while asking the question. Using the example illustrated in Table 15, the implied numerosity framework suggests that unpacking England and Germany in response to the question "which country saw the invention of the bicycle" should lead the judge to focus on a category roughly the size of Europe because all explicitly presented hypotheses are European countries. So, providing the hint that the correct response is a European country should have no effect on the probability assigned. The similarity in responses in the SU and LUS conditions, therefore support the idea that the alternatives presented led judges to focus on a category of sufficient size to encompass all explicitly mentioned alternatives and nothing more. In addition to this being viewed as a manipulation check, the similar ratings in these two conditions also rule out the possibility that the higher probability assigned to the focal diagnoses in the Small Packed condition relative to the Large Packed condition resulted solely from an increase in confidence as a result of having been given a hint.

Insert Table 17 About Here

Discussion

Like Cosmides and Tooby (1995), I propose that humans may be good intuitive statisticians after all. It is a rational act to assign a probability on the basis of the number of potential alternatives -- all else being equal, the more alternatives there are, the lower the likelihood that any one alternative will be correct. Unlike Cosmides and Tooby, I do not think that the phenomena observed in much of the judgment under uncertainty literature is simply a numbers game that results from a poor conceptualization of probabilities. Rather, I argue that systematic variations found in probability judgments arise, in part, as a result of conceptions being altered by the way in which questions are asked or by the number of alternatives that one is explicitly asked to evaluate. The presentation of specific alternatives not only influences the support in favour of those alternatives, but can also drive people's impressions of the number of potential alternatives that they have not considered explicitly. A physician who is asked to evaluate a medical case for the presence of two specific infectious disorders might perceive the problem differently than one who is asked to evaluate the same problem via the consideration of an infectious disorder and a genetic defect. Even if additional alternatives are not explicitly considered, the two physicians' senses as to the number of plausible alternative diagnoses that exist is likely very different, but nonetheless potentially influential in both cases.

Again, the argument is not that implied numerosity is the only factor influencing probability judgments. On the contrary, Support Theorists have shown that a more explicit consideration of specific alternatives is very influential. Rather, I speculate that judges do

not consciously evaluate the number of alternatives that are plausible before transforming the result into a probability judgment or frequency estimation, and yet an implicit appreciation of the number of alternatives affects the probability judgment nonetheless. There remain at least two possible mechanisms through which judges might be sensitive to implied numerosity.

It has been argued throughout this chapter that the alternatives explicitly mentioned create a context through which judges gain a general impression of the number of possible alternatives that exist. A second possibility is that the explicit mention of examples that constitute a relatively large category cue more additional alternatives than do examples that constitute a smaller category. That is, the unpacked hypotheses might provide an explicit numerosity by overtly reminding individuals of specific alternative hypotheses. The latter possibility seems unlikely for two reasons. First, the phenomenology of working through a probability judgment is rarely one of alternative hypothesis generation. As mentioned earlier, participants in these studies rarely, if ever, report expending a significant amount of effort spontaneously generating alternatives. Second, researchers examining concept formation (e.g., McRae, de Sa, and Seidenberg, 1997) often opt to use larger category decision tasks (e.g., superordinate descriptors such as "is it a living thing") so as to avoid cueing specific exemplars. Jared and Seidenberg (1992) found that narrower decision tasks (e.g., "is it a bird") were inappropriate, because they were more likely to cue specific exemplars. If a similar phenomenon existed during the formulation of probability judgments, then a greater amount of spontaneous unpacking would have been expected in the "small category" condition relative to the "large category" condition, thereby resulting in predictions opposite to the ones confirmed in this chapter. In summary, it appears unlikely that frequency estimations consist solely of a systematic search through multiple possible alternatives. Rather, I argue that information is gained through implied numerosity,

which consists of the creation of a general impression of the size of the category under consideration, and which thereby modulates the judged probability.

CHAPTER 8

General Discussion

Summary

When assessing the probability of a particular event, an individual's judgment is influenced by the context within which the judgment is made. Judges tend to under-appreciate the probability of diagnoses that remain implicit components of the residual category relative to when they are presented for explicit evaluation. This is true even when the judges are diagnosticians who are regularly admonished during training to carefully consider differential diagnoses before concluding a diagnostic search. In fact, Experiment 3 revealed that diagnoses that participants generate and claim to have taken into account while making their judgment are under-weighted relative to when the same diagnoses are presented explicitly. As a result, clinical instruction to generate differential diagnoses is likely to be insufficient to eliminate the under-weighting bias illustrated by the unpacking effect.

Experiments 4 to 6 provide evidence consistent with the hypothesis that the mechanism responsible for the under-weighting of implicit diagnoses is differential processing of the available evidence at the point of decision. The presence of a diagnostic suggestion appears to direct one's focus of attention preferentially towards features of the case that are consistent with the focal diagnosis relative to those features that are consistent with non-focal alternatives. Anecdotal evidence and Experiment 7 suggest that particularly salient features tend to capture attention when individuals are not careful to consider the rest of the evidence during the decision process. Instruction to re-evaluate the evidence within a

case before assigning a probability assessment better enabled participants to weight the non-focal alternatives in a consistent manner. This was evidenced by the elimination of the unpacking effect. It is important to note that re-evaluation of the evidence at the point of decision is critical. One can not avoid under-weighting implicit diagnoses by simply undertaking a more controlled evaluation of the evidence in the case. Experiment 6 illustrates that it is not enough simply to collect the data more carefully, but that the information collected must be considered during the course of making the necessary judgment.

Furthermore, instruction to re-evaluate the evidence before making a decision does not necessarily depend on explicitly redirecting attention to a specific non-focal diagnosis. In Experiment 5, instruction to list the evidence that was consistent with the focal diagnosis appears to have been sufficient to get the student diagnosticians to realize that evidence was present in the case that was not supportive of the focal diagnosis.

Another way to avoid under-weighting non-focal diagnoses is to allow diagnosticians to generate the focal hypotheses for themselves. In the absence of a diagnostic suggestion, one cue that can cause the available evidence to be evaluated differentially is removed. This does not mean that the biases that do arise when an explicit diagnostic suggestion is presented are unimportant. Diagnosticians, especially students, often engage a clinical case with a preconceived probable diagnosis. In fact, diagnostic suggestions provide an effective way of directing clinical teaching. Furthermore, as the use of computer aided diagnostic systems is rapidly proliferating, the influence of explicit diagnostic suggestions will become even more of an issue. Medical educators must, therefore, be aware that (a) diagnostic suggestions are likely to bias a student's conception of the case and (b) simple instruction to consider differential diagnoses is insufficient to enable students to avoid premature closure.

Finally, it should be stressed that I am not arguing that differential processing of the evidence available is the only mechanism that leads to under-weighting of the residual "all other diagnoses" category. On the contrary, the results of Experiment 9 show that weighting of support for the non-specific residual can also be biased by the type of diagnoses that are explicitly presented. The explicit presentation of a subset of alternatives appears capable of altering a judge's perception of the number of alternatives that are "out there," but not brought to mind immediately. This suggests that a judge will under-weight the residual category to a greater extent if the diagnoses presented distract the judge from realizing that additional, but inaccessible, diagnoses exist. That is, the numerosity of alternatives that is implied by the specific diagnoses considered will influence the extent to which an individual is biased by the explicitly presented focal diagnosis.

Implications

The experiments presented in this dissertation advance our understanding of the way in which humans assess the probability of events. Previous work by support theorists and others have nicely illustrated that our judgments are not logically consistent with the standards of probability theory. The current dissertation, I believe, helps to further understand the mechanisms that produce such biases. The results herein strongly implicate attentional processes as influential in the way in which the context of the question to be answered is construed.

While informing psychological theory about how probability judgments are made, these studies also have practical implications. Although it is hoped that the results generalize to all decision-making contexts, including consumerism, I have chosen to focus primarily on diagnostic decision-making. As medical educators strive to teach students not to be fooled into prematurely closing their diagnostic search, the current data suggest that

simply instructing students to think of differential diagnoses will be insufficient. When asking students to comment on or evaluate the probability of a particular diagnosis, educators need to encourage students to take a step back from the diagnosis and re-evaluate the evidence that is available. Failure to do so is likely to result in attention being focused primarily on the evidence consistent with the focal diagnosis, which will in turn bias the responses provided. Although such instruction will not perfect medical students' diagnostic decisions, they may eliminate one persistent source of error in those decisions.

Using our understanding of these biases to avoid errors in diagnosis is likely to become more important as emphasis on the use of Evidence-Based diagnostic strategies increases. As Fletcher, Fletcher, and Wagner have noted, "increasingly, the modern clinician expresses the likelihood that a patient has a disease by using a probability" (1996, p.44). Practicing Evidence-Based Medicine requires that physicians assign probability ratings and adjust these ratings up or down as more light is shed on the case (Sackett, Haynes, Guyatt, and Tugwell, 1991). Although such attempts to objectify the evidence at hand should be applauded, an obvious constraint on the efficacy of these techniques is the accuracy of the subjective probability judgment that serves as the starting point.

Similarly, our understanding of the influence of explicit diagnostic suggestions might also inform medical practice that uses computer aided diagnostic systems. These software packages typically operate by generating a list of plausible diagnoses in response to input from the physician regarding which signs and symptoms have been detected. Assuming for the moment that input can be provided objectively, the output provided should enable a final diagnosis that is not biased by a specific focal hypothesis. The work outlined in this dissertation suggests that these systems need to be implemented in a way that allows the computer the opportunity to make non-focal alternative diagnoses explicit and, in turn, helps the diagnostician avoid differentially processing the evidence available. However, Brooks, LeBlanc, and Norman (2000) raise concern about a self-fulfilling prophecy that

might be created when using diagnostic software packages that require the input of diagnostician-detected features. The problem is that symptoms that are consistent with a focal diagnosis are more likely to be detected than are symptoms consistent with a non-focal diagnosis. Whether this problem can be avoided remains to be seen.

Limitations

There are some potential problems with the work presented within this dissertation that might limit the extent to which the results can be generalized. First, the use of medical students as participants might lead one to question whether the results observed can be generalized to a more expert population. Perhaps medical personnel are able to overcome the biases observed as their diagnostic expertise increases, thereby limiting the applicability of these results to novice diagnosticians. Two observations mitigate this concern. First, to some extent it does not matter whether the effects described here are limited to the decision-making abilities of novices. If the present results apply only to novice decisions, there are still useful applications to improving clinical teaching, thereby enabling novices to learn the art of diagnosis more efficiently. However, this caution might be unwarranted, because the results described here very likely will generalize to expert diagnosticians. Norman, Coblentz, Brooks, and Babcock (1992) reviewed a series of studies examining expertise in visual diagnosis and concluded that experts and novices do not differ in the process used to make a diagnosis – experts simply perform the process better. Furthermore, many biasing studies have confirmed that both experts and novices tend to be susceptible to the same heuristic-induced errors (Redelmeier, et al., 1995; Green and Yates, 1995; Hatala, Norman, and Brooks, 1999). However, it is true that experts, almost by definition, are more likely than novices to evaluate a case within the context of the correct diagnosis. As a result, perhaps experts are as susceptible as novices to differentially focusing their attention on the

evidence, but such biased processing results in a diagnostic error less often simply because the focal diagnosis is more likely to be correct. Although the number of errors a diagnostician makes undoubtedly decreases with the development of expertise, understanding the heuristics that are used can potentially allow insight into the source of the few errors that are made.

A second limitation arises from the possibility that my participants did not approach the paper and pencil clinical cases that I provided as they normally would a real case. However, it is at least as likely that paper and pencil cases underestimate the strength of the biases inherent in diagnostic decision-making. There was no opportunity or need for participants to direct the collection of data. As a result, all participants in my studies were assured to receive all of the necessary information for both focal and non-focal disorders. Providing subjects with control over the collection of data could make the effects stronger, as the bias towards considering the evidence consistent with the focal diagnosis might cause participants to collect only that information. This bias in evidence collection would reduce the amount of evidence available that is consistent with non-focal diagnostic alternatives in addition to creating a tendency to under-weight the information that is collected. This is an important empirical question that has not yet been tested.

An additional limitation might be that participants may not have exerted as much effort in working through these cases as they typically do in everyday medical situations. Economists argue that the only way to ensure that individuals make judgments that they truly believe in is to include consequences for poor judgments (e.g., loss of financial gain). Although this might be true, medical participants tend to be very motivated individuals who have reason to be concerned about their ability to diagnose. There were no physical consequences in terms of effects on patients, but poor decisions do carry the consequence of bruised egos and concern regarding one's diagnostic ability. Consistent with this argument is the fact that many of the participants took the study seriously enough that they

requested the opportunity to confer about the cases within their tutorial groups after completion of the study.

Finally, the use of probability judgments as the primary dependent variable might raise concern. It is possible that changes in the probability assigned to a particular diagnosis do not mimic changes in the actual decisions that are reached by diagnosticians. However, probability judgments have been shown by Redelmeier et al. (1995) to correlate with more ecologically valid measures of the decision process (i.e., the type of diagnostic tests that physicians would like to have performed). Experiment 3 revealed a similar result in that the number of diagnostic tests ordered tends to be greater when a greater number of diagnostic hypotheses are explicitly presented. This change in patient management strategies supports my assumption that the assignment of probability ratings provides an adequate measure of the way in which a clinical case is being perceived. Furthermore, the growing emphasis that is being placed on the use of evidence-based medicine means that probability judgments themselves are becoming increasingly ecologically valid measures of the diagnostic decision-making process.

Future Directions

There are a number of questions that the current work leaves unanswered, many of which have been raised in the context of particular experiments. Rather than repeating each discussion, I will simply elaborate on one particularly interesting possibility. It was argued in Chapter 7 that both analytic (i.e., the consideration of specific diagnostic hypotheses) and non-analytic (i.e., implied numerosity) processes might contribute to a diagnostician's probability assessments. It might be particularly informative to know whether the relative contribution of each type of process is variable across situations. For example, it is possible that, as experts develop a greater database of prior experiences they are better able to avoid

generating a "laundry list" of specific diagnostic possibilities, instead relying to a greater extent on non-analytic assessments. If this turned out to be true, then more specific statements could be made regarding how a diagnostician might avoid making biased responses across different contexts. For now, however, it will have to suffice to say that clinical instruction to generate differential diagnoses does not eliminate premature closure. The presence of an explicitly mentioned diagnostic hypothesis can result in just fleeting consideration of self-generated hypotheses. So, to better enable medical students to avoid under-weighting non-focal diagnoses, students should be instructed to re-evaluate the evidence available at the point at which the final judgment is made.

REFERENCES

Anderson, R.C. and Pichert, J. (1978). Recall of previously unrecallable information following a shift in perspective. Journal of Verbal Learning and Behavior, 17, 1-12.

Arkes, H.R., Saville, P.D., Wortmann, R.L., and Harkness, A.R. (1981). Hindsight bias among physicians weighing the likelihood of diagnoses. Journal of Applied Psychology, 66, 252-254.

Ayton, P. (1997). How to be incoherent and seductive: Bookmakers' odds and support theory. Organizational Behavior and Human Decision Processes, 72, 99-115.

Begg, I.M., Faulkner, H.J. and Jacoby, L.L. (Unpublished manuscript). Dissociation of processes in frequency discrimination: The frequency attribute is controlled and intentional but the availability heuristic is automatic.

Bordage, G, and Lemieux, M. Semantic structures and diagnostic thinking of experts and novices. Academic Medicine, 66 (Suppl.), S70-S72.

Brenner, L.A., and Koehler, D.J. (1999). Subjective Probability of Disjunctive Hypotheses: Local-Weight Models for Decomposition of Evidential Support. Cognitive Psychology, 38, 16-47.

Brooks, L.R., LeBlanc, V.R. and Norman, G.R. (2000). On the difficulty of noticing obvious features in patient appearance. Psychological Science, 11, 112-117.

Cosmides, L., and Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions of the literature on judgment under uncertainty. Cognition, 58, 1-73.

Cunnington, J.P.W., Turnbull, J.M., Regehr, G, Marriott, M and Norman, G.R. (1997). The effect of presentation order in clinical decision making. Academic Medicine, 72 (10 Suppl. 1), S40-S42.

Duffy, J.E. (1994). Modern Automotive Technology. South Holland, Illinois: The Goodheart-Wilcox Company, Inc.

Ericsson, K.A, and Charness, N. (1999). Expert performance: Its structure and acquisition. In The nature-nurture debate: The essential readings. Essential readings in developmental psychology. S.J. Ceci, W.M.Williams et al. (Eds.). Malden, MA: Blackwell Publishers.

Fischoff,B., Slovic, P., and Lichtenstein, S. (1978). Fault trees: Sensitivity of estimated failure probabilities to problem representation. Journal of Experimental Psychology: Human Perception and Performance, 4, 330-344.

Fletcher, R., Fletcher, S., and Wagner, E. (1996). Clinical Epidemiology: The Essentials. (2nd ed). Baltimore: Williams & Wilkins.

Green, L.A. and Yates, J.F. (1995). Influence of pseudodiagnostic information on the evaluation of ischemic heart disease. Annals of Emergency Medicine, 25, 451-457.

Hatala, R, Norman, G.R. and Brooks, L.R. (1999). The impact of a clinical scenario upon accuracy of electrocardiogram interpretation. Journal of General Internal Medicine, 14, 126-129.

Hawkins, S.A., and Hastie, R. (1990). Hindsight: Biased judgments of past events after the outcomes are known. Psychological Bulletin, 107, 311-327.

Jared, D., and Seidenberg, M.S. (1992). Does word identification proceed from spelling to sound to meaning? Journal of Experimental Psychology: General, 120, 358-394.

Johnson, E.J., Hershey, J., Meszaros, J., and Kunreuther, H. (1993). Framing, probability distortions, and insurance decisions. Journal of Risk and Uncertainty, 7, 35-51.

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

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

Koehler, D.J., Brenner, L.A., and Tversky, A. (1997). The Enhancement Effect in Probability Judgment. Journal of Behavioral Decision Making, 10, 293-313.

Krenz, I., Wolf, G., and Stahl, R.A. (2000). Premature closure or do not get lost in your diagnostic work-up and blame it on the patient. Nephrology, Dialysis, and Transplants, 15, 1072-1075.

Lingard, L.A., and Haber, R.J. (1999). What do we mean by "relevance?" A clinical and rhetorical definition with implications for teaching and learning the case-presentation format. Academic Medicine, 74 (10 Suppl), S124-S127.

McKenzie, C.R.M. (1998). Taking into account the strength of an alternative hypothesis. Journal of Experimental Psychology: Learning, Memory, and Cognition, 24, 771-792.

McRae, K., de Sa, V.R., and Seidenberg, M.S. (1997). On the nature and scope of featural representations of word meaning. Journal of Experimental Psychology: General, 126, 99-130.

Morris, C.D., Bransford, J.D., & Franks, J.J. (1977). Levels of processing versus transfer appropriate processing. Journal of Verbal Learning and Verbal Behavior, 16, 519-533.

Mulford, M., and Dawes, R.M. (1999). Subadditivity in memory for personal events. Psychological Science, 10, 47-51.

Norman, G.R., Coblentz, C.L., Brooks, L.R., and Babcock, C.J. (1992). Expertise in visual diagnosis: a review of the literature. Academic Medicine, 67(10 Suppl), S78-83.

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

Reisberg, D. (1997). Cognition: Exploring the science of the mind. New York: W.W. Norton & Company.

Rottenstreich, Y., and Tversky, A. (1997). Unpacking, repacking, and anchoring: Advances in Support Theory. Psychological Review, 104, 406-415.

Sackett, D.L., Haynes, R.B., Guyatt, G.H., and Tugwell, P. (1991). Clinical Epidemiology: A Basic Science for Clinical Medicine. (2nd ed). Toronto: Little Brown and Company.

Teigen, K.H. (1982). Studies in subjective probability III: The unimportance of alternatives. Scandinavian Journal of Psychology, 24, 97-105.

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson (Eds.), Organization of memory. New York: Academic Press.

Tversky, A., and Fox, C.R. (1995). Weighing risk and uncertainty. Psychological Review, 102, 269-283.

Tversky, A., and Kahneman, D.J. (1983). Extensional vs. intuitive reasoning: The conjunction fallacy in probability judgment. Psychological Review, 91, 293-315.

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

Voytovich, A.E., Rippey, R.M., and Suffredini, A. (1985). Premature conclusions in diagnostic reasoning. Journal of Medical Education, 60, 302-307.

Wolf, F.M., Gruppen, L.D. and Billi, J.E. (1985). Differential diagnosis and the competing-hypotheses heuristic: A practical approach to judgment under uncertainty and Bayesian probability. JAMA, 253, 2858-2862.

Table 1

Hypothesis list presented for fibromyalgia case, as a function of condition (Experiment 1)

1A	$5A_C$	$5A_U$	$5A_I$
Fibromyalgia	Depression	Lupus	Polymyalgia rheumatica
All other diagnoses	Fibromyalgia	Rheumatoid arthritis	Myeloma
	Chronic fatigue syndrome	Fibromyalgia	Endocarditis
	Anemia	Polymyositis	Addison's disease
	Hypothyroidism	Renal failure	Fibromyalgia
	All other diagnoses	All other diagnoses	All other diagnoses

Table 2

Mean probability ratings (expressed as percentages) assigned to non-focal diagnoses by clinical clerks, as a function of type of condition (Experiment 1)

	5A (Implausible)	5A (Uncommon)	5A (Common)
Most Likely Alternative Diagnosis	11	11	22
Sum (All Alternative Diagnoses)	22	29	50

Table 3

Mean probability ratings (expressed as percentages) assigned to focal diagnoses by clinical clerks, as a function of type of condition (Experiment 1)

Condition	1 Alternative	5 Alternatives		
		Implausible	Uncommon	Common
First	40	37	33	25
Mixed	42	35	34	25
Combined	41	36	33	25
Standard Deviation	(16)	(17)	(17)	(12)

Table 4

Mean probability ratings (expressed as percentages) assigned to focal diagnoses by motive technology students, as a function of type of condition and expertise (Experiment 2)

Expertise Level	1 Alternative	Implausible	3 Alternatives Uncommon	Common
Common Core	49.2	36.0	30.8	22.8
Basic	42.1	38.9	34.0	21.1
Advanced	42.1	32.6	28.2	22.6
All Levels Combined	44.4	35.4	29.9	22.3
Standard Deviation	19.0	20.8	19.0	16.0

Table 5

Mean probability ratings, expressed as percentages (and count of the number of observations), assigned to the focal diagnoses by medical students, as a function of session, the number of diagnoses presented during session 1, and whether or not an alternative diagnosis was generated by the student during session 1 (Experiment 3)

Session	Diagnosis(es) Presented	Focal Diagnosis if Alternative Generated	Focal Diagnosis if Alternative Not Generated	Overall
1	Focal	44.39 (89)	60.83 (6)	45.43 (95)
	Focal + Alternative	33.86 (70)	38.80 (25)	35.16 (95)
	Overall	39.75 (159)	43.06 (31)	40.29 (190)
2	Focal (+ Generated Alternative if Generated)	37.13 (89)	59.17 (6)	38.53 (95)
	Focal + Alternative (+ Generated Alternative if Generated)	30.07 (70)	42.00 (25)	33.21 (95)
	Overall	34.03 (159)	45.32 (31)	35.87 (190)

Table 6

Example of a patient profile (Experiment 4)

Symptom	Present?
Dizziness	Yes
Rash	No
Fever	No
Coughing	Yes

Table 7

Symptom by disease matrix illustrating categorization scheme for puneria and zymosis (Experiment 4)

	Puneria	
Symptom	P (symptom | puneria)	P (symptom | normal)
Dizziness	.80	.20
Rash	.60	.20
Coughing	.40	.20
Fever	.20	.20
	Zymosis	
Symptom	P (symptom | zymosis)	P (symptom | normal)
Dizziness	.20	.20
Rash	.40	.20
Coughing	.60	.20
Fever	.80	.20

Table 8

Mean absolute deviation (and standard deviations) from 100 for confidence ratings summed across both illnesses for all 16 test patient profiles, as a function of question type (Experiment 4)

	Asymmetric	Symmetric	Evidence
Mean	39.3	23.0	13.9
Standard Deviation	23.97	11.76	9.90

Table 9

Mean probability rating (expressed as percentages) assigned to the focal diagnosis, as a function of the type of residual category, and instruction (Experiment 5)

	Immediate	Argue for F	Argue for F & NF
Packed	44.8	38.9	47.3
Unpacked	33.8	35.3	45.3
Unpacking Effect (Packed – Unpacked)	11.0 ($p < .05$)	3.6 ($p > .5$)	2.0 ($p > 0.5$)

Table 10

Mean probability rating (expressed as percentages) assigned to the focal diagnosis, as a function of the type of residual category, and instruction (Experiment 6)

	Delay	Distracting	Immediate
Packed	55.8	47.8	40.7
Unpacked	41.6	41.9	36.9
Unpacking Effect (Packed – Unpacked)	14.2 ($p < .05$)	5.9 ($p > .1$)	3.8 ($p > 0.25$)

Table 11

Description of manipulated features for each case (Experiment 7)

Note: These features are just a subset of those present in a longer (1/2 page single spaced) case history

Case	Diagnosis	Feature	
		Medicalese Description	Lay Description
1.	Pulmonary Embolus	- The dyspnea occurred suddenly at 11:00 pm and had awoken the patient from sleep	- The shortness of breath occurred at 11:00 pm and had awoken the patient from sleep
		- She complained of retrosternal chest pain that was worse on deep breathing	- She complained of chest pain that was worse on deep breathing
		- For several days she had also experienced hemoptysis	- For several days she had also coughed up blood
	Pneumonia	- For 4 days she felt unwell and had a sore throat and sinus congestion that had resolved	- For 4 days she felt unwell and had a sore throat and stuffiness that had resolved
		- She complained of fever and chills during several occasions in the past few days.	- She complained of feeling hot and cold during several occasions in the past few days
		- Her WBC count was elevated (12,000), but her Hemoglobin levels were normal (132)	- The WBC is 12,000 and Hb 132

(cont'd on next page)

2.	Bronchitis	- For several years he has had a chronic productive cough	- For several years he has had a cough
		- On several occasions in the past few days he has experienced episodes of fever and chills ...	- On several occasions in the past few days he has experienced feeling hot and cold.
		- ... there have been no rigors	- Otherwise he has felt normal
	Lung Cancer	- The hemoptysis continued for several days.	- [Coughing up bright red blood] continued for several days
		- He has been trying to lose weight and over the past 2 months he has experienced significant weight loss (6 pounds)	- He has been trying to lose weight and over the past 2 months he has lost 6 pounds
		- There has been no wheezing and no hoarseness	- Breathing has been normal
3.	Gastritis	- Three years previously he had been treated at a walk-in clinic with a one-month prescription for a possible ulcer	- Three years previously he had been treated at a walk-in clinic with a one-month prescription for stomach pain
		- He suffers from occasional heartburn	- He occasionally experiences warmth in his chest after eating
		- There is no family history of heart disease	- His family history is unremarkable
	Myocardial Ischemia	- There has been some dyspnea	- There has been some shortness of breath
		- A year ago, during an executive physical, his cholesterol was at the upper limits of normal (13 mmol/L)	- A year ago, during an executive physical, his cholesterol was 13 mmol/L
		- He smokes a substantial number of cigarettes per day, but has been trying to cut down	- He smokes a pack of cigarettes per day, but has been trying to cut down

(cont'd on next page)

4.	Diuretic Induced	- There is a five year history of hypertension for which she takes 50 mg. of hydrochlorthiazide daily	- There is a five year history of elevated blood pressure for which she takes 50 mg. of hydrochlorthiazide daily
		- There is no clubbing or cyanosis	- The rest of the exam is normal
		- There is no respiratory distress	- The rest of the exam is normal
	SIADH	- For the past several years she suffered from dyspnea, particularly on exertion	- For the past several years she suffered from shortness of breath particularly when doing her food shopping
		- She has had a chronic productive cough for several years and it is not clear if it has recently been worse	- She has had a cough for several years and it is not clear if it has recently been worse
		- She has lost a substantial amount of weight in the past year (5-10 pounds), but her appetite has always been poor	- She has lost 5-10 pounds in the past year, but her appetite has always been poor
5.	Inflammatory bowel disease	- His diarrhea has been prolonged (10-14 days before presentation) and is worsening in severity	- His diarrhea began 10-14 days before presentation and is worsening in severity
		- Review of systems revealed some mild arthralgias ...	- Review of systems revealed some joint pain
		- ... and some nausea and vomiting	- He also reports queasiness and feeling sick to his stomach
	Infectious Disorder	- On examination he has a fever (37.8 C)	- On examination he has a temperature of 37.8 C
		- The locomotor exam is normal with no evidence of arthritis	- The rest of the exam is unremarkable
		- There is no skin rash	- The rest of the exam is unremarkable

(continued on next page)

6.	Gall Stones	- He abdominal pain is dull and episodic	- Her abdominal pain comes and goes and does not feel sharp
		- The pain gets worse after eating and does not radiate	- The pain gets worse after eating. Otherwise it's unremarkable
		- Abdominal examination reveals mild tenderness in the epigastrium	- Abdominal examination reveals tenderness near the stomach
	Peptic Ulcer Dx.	- There has been no jaundice ...	- The remainder of the examination is normal
		- ... or fever	- The remainder of the examination is normal
		- On examination she is an obese female	- She weighs 190 pounds

Table 12

Mean probability rating (expressed as percentages) assigned to each diagnosis, as a function of case and terminology (Experiment 7)

Panel a) Case number 1

	Diagnosis	
Version	A: Pulmonary Embolus	B: Pneumonia
A Medicalese, B Lay	63.57	17.86
B Medicalese, A Lay	57.14	26.43

Panel b) Case number 2

	Diagnosis	
Version	A: Bronchitis	B: Lung Cancer
A Medicalese, B Lay	52.86	32.14
B Medicalese, A Lay	31.43	52.86

Panel c) Case number 3

	Diagnosis	
Version	A: Gastritis	B: Myocardial Ischemia
A Medicalese, B Lay	51.43	28.14
B Medicalese, A Lay	62.86	24.29

(continued on next page)

Panel d) Case number 4

	Diagnosis	
Version	A: Diuretic Induced	B: SIADH
A Medicalese, B Lay	55.00	21.43
B Medicalese, A Lay	42.14	37.29

Panel e) Case number 5

	Diagnosis	
Version	A: Inflammatory Bowel Disease	B: Infection
A Medicalese, B Lay	67.86	27.71
B Medicalese, A Lay	55.71	29.29

Panel f) Case number 6

	Diagnosis	
Version	A: Peptic Ulcer	B: Gall Stones
A Medicalese, B Lay	51.86	29.71
B Medicalese, A Lay	50.71	35.71

Table 13

Mean number of features recalled (and % of total number possible), as a function of terminology and whether or not feature was manipulated (Experiment 7)

	Manipulated Feature	Non-manipulated Feature	Total
Medicalese	189 / 504 (37.5%)	128 / 248 (51.7%)	417 / 752 (55.5%)
Lay Terminology	125 / 504 (24.8%)	104 / 248 (41.9%)	249 / 752 (33.1%)

Table 14

Mean probability ratings, expressed as percentages (and count of the number of observations), assigned to the focal diagnoses by medical students, as a function of session, and whether or not an alternative diagnosis was generated by the student during session 1 (Experiment 8)

Session	Diagnosis(es) Presented	Focal Diagnosis if Alternative Generated	Focal Diagnosis if Alternative Not Generated	Overall
1	Focal	66.1 (227)	63.3 (19)	65.8 (246)
2	Focal (+ Generated Alternative if Generated)	71.3 (227)	69.1 (19)	71.1 (246)

Table 15

Example question illustrating five conditions (Experiment 9)

"Which country saw the invention of the bicycle?"

Condition	Hint	Alternatives
Large Packed		France, "All other countries"
Small Packed	It's in Europe	France, "All other European countries"
Large Unpacked Similar		France, England, Germany, "All other countries"
Large Unpacked Dissimilar		France, U.S.A., Taiwan, "All other countries"
Small Unpacked	It's in Europe	France, England, Germany, "All other European countries"

Table 16

Mean probability rating (expressed as percentages) and standard deviations assigned to the focal diagnosis, as a function of the type of residual category (Experiment 9)

	Large Category		
	Packed	Unpacked Similar	Unpacked Dissimilar
Mean	44.5	32.6	28.7
Standard Deviation	17.21	15.26	14.55

Table 17

Mean probability rating (expressed as percentages) and standard deviations assigned to the focal diagnosis, as a function of the type of residual category (Experiment 9)

	Large Category	Small Category
Packed	44.5 (17.21)	49.2 (17.77)
Unpacked Similar	32.6 (15.26)	30.9 (13.15)
Unpacking Effect (Packed – Unpacked)	11.9 (18.03)	18.3 (19.10)