Cognition, 49 (1993) 97–122
0010-0277/93/$06.00

Debias the environment instead of the judge: an alternative approach to reducing error in diagnostic (and other) judgment

Joshua Klayman*[,a], Kaye Brown[b]
[a]*Center for Decision Research, Graduate School of Business, University of Chicago, 1101 East 58th Street, Chicago, Illinois 60637, USA*
[b]*National Center for Health Program Evaluation, Fairfield, Victoria, Australia*

Abstract

Questions about how to improve human judgment and reasoning are of theoretical and practical interest, notwithstanding the continuing controversy over whether people are "rational". Improving judgment may involve modifying people's processes to fit their environments better, or vice versa. We illustrate the latter approach in a study of diagnostic reasoning in which subjects learned to distinguish two fictitious diseases. Prior findings suggest that people may judge the likelihood of a diagnostic category on the presence or absence of features that are typical of, rather than diagnostic of, the category. We varied the structure of the information provided to subjects without attempting to modify their judgmental processes. In an "independent" format, subjects learned about each disease separately; in a "contrastive" format, information about the two diseases was juxtaposed to highlight distinctive features. Subjects in the two conditions formed different disease concepts. Diagnoses following contrastive training were much closer to the statistically prescribed judgments based on likelihood ratios. Interventions that modify the environment may provide an alternative approach where it is difficult to modify people's processes. Effective design of such interventions is one motivation

We thank Jackie Gnepp, Arthur Elstein, Michael Doherty, Craig McKenzie, F.R. Heath, Linda Ginzel, Selwyn Becker, Robin Hogarth, and an anonymous reviewer for their valuable comments on earlier drafts of this paper.
*Corresponding author.

for directing research toward understanding how task characteristics affect the use of and the outcomes of judgment and reasoning processes.

Introduction

What is wrong with the way people think, and what can be done to improve it? A large proportion of the research on human judgment, reasoning, and decision making has been concerned with variations on these basic themes. Controversy continues not only over the answers, but also over the legitimacy of the questions themselves: is there anything systematically wrong with human thinking and if, in fact, it *ain't* broke, how can you fix it?

We will sidestep the debate over whether human thinking is fundamentally rational or irrational (see Abelson & Levi, 1985; Cohen, 1981; Edwards, 1990; Funder, 1987; Hammond, 1990; Hogarth, 1981; Jungerman, 1983; Kyburg, 1983), and the related question posed by Anderson (1990, 1991), "is human cognition adaptive?" We take an agnostic position: human judgment may be well *adapted*, but it is not always very *adaptable*. That is, cognitive processes may be well matched to the requirements of the average, modal, or typical task to which the process is applied, and yet not be ideally suited to the demands of the particular task at hand (e.g., see McKenzie, in press). Thus, error in judgment and reasoning can be viewed as indicative of mismatches between the cognitive processes people use and the tasks to which those processes are applied (see also Fischhoff, 1982). This view is not incompatible with the possibility that human cognitive processes are well suited to the general environment. Nevertheless, there may be identifiable subenvironments in which people could be doing better given their goals and their resources.

Modifying cognitive processes to fit the task

The view that problems come from discrepancies between tasks and processes suggests a taxonomy for interventions, based on how they attempt to bring processes and tasks into better alignment. A variety of approaches have been proposed to help change people's cognitive processes to fit their tasks better (see Fischhoff, 1982; Kahneman & Tversky, 1979; von Winterfeldt & Edwards, 1986):

Teach people more adaptable processes. One approach is to provide people with better cognitive tools. This may include training in general methods (e.g., statistical approaches to variance and sampling) that can replace intuitive heuristics (e.g., representativeness) in appropriate situations (Beyth-Marom,

Dekel, Gombo, & Shaked, 1985; Fong, Krantz, & Nisbett, 1986). Or, people may be taught to generalize useful "statistical heuristics" they have learned in specific domains (Nisbett, Krantz, Jepson, & Kunda, 1983).

Teach people to anticipate and correct their mistakes. Interventions can also focus on correcting, rather than avoiding, mistakes. For example, Koriat, Lichtenstein, and Fischhoff (1980) had subjects counter their tendency toward overconfidence in a chosen answer by explicitly listing reasons for and against their choice. Gaeth and Shanteau (1984) also found that correction was more feasible than avoidance in training people not to be influenced by irrelevant information in multiple-cue judgment tasks such as evaluating soil samples or job candidates. They found that people could not learn to ignore cues that were normatively irrelevant, but could learn to recognize the effects those cues had on their judgments, and to compensate for them.

Train people in task-specific processes. In some circumstances, it may be more feasible to teach not general problem-solving skills, but domain-specific heuristics. Such heuristics might be *less* well adapted in the overall environment, but would be better in a specific subenvironment. Smith and Kida (1991), for example, found that professional auditors had developed domain-specific strategies for testing and revising hypotheses (e.g., paying extra attention to any evidence of trouble, whether anticipated or not), and Dunbar (in press) observed that researchers in molecular biology learned, through observation, direct instruction, and painful experience, to distrust exciting positive findings. For both the auditors and the scientists, domain-specific training reduced the incidence of biases related to perseverance of beliefs. How well such training generalizes is unclear. How do Dunbar's scientists behave when developing ideas about child rearing, for example? Conventional wisdom in the expertise literature suggests transfer beyond the domain of training will be limited. Smith and Kida (1991) found, for example, that confirmatory biases in hypothesis testing and biases associated with the representativeness heuristic reappeared nearly full force if the experts were asked to make judgments they were not practiced at making, even when those judgments were relevant to their professional activities. On the other hand, Lee and Pennington (1992) report that expertise in debugging computer programs generalized well to unfamiliar problems in troubleshooting electronic circuits.

Provide people with analytical tools. Another widely used approach to improving judgment is to develop analytical techniques that people can use to supplement or supplant their intuitions. Examples include multi-attribute utility analysis, linear modeling, and Bayesian statistics. There is little doubt that people can be trained to use such techniques effectively to make better decisions under many circumstances. However, formal analysis is time consuming, effortful, and often difficult to do correctly (see von Winterfeldt & Edwards, 1986). For problems in which time, effort, and information are constrained, various

hybrids of analytical and intuitive approaches may be more appropriate (see Hammond, Hamm, Grassia, & Pearson, 1987; Schoemaker & Russo, 1992).

All of these approaches hold promise for improving human judgment for all of the people some of the time. However, there is another, qualitatively different approach to improving judgments that has received relatively little attention. Rather than modifying cognitive processes to fit the environment better, one can modify the environment to fit the processes that people bring to it.

Modifying the environment to fit cognitive processes

In human factors and industrial design, one often thinks about modifying environments to fit people, as well as vice versa. If your operators have trouble reading their instruments, you might think about providing an easier-to-read instrument panel or making the operating manual easier to follow. Analogously, the knowledge that has accumulated concerning human judgment and reasoning processes could be used to design environments that avoid or compensate for anticipated errors.

One way to improve performance is to make the required processes easier to execute. In some cases, people may already be capable of more veridical processes, but are constrained by the costs of implementing them properly. Improving judgment in such cases may merely require designing an environment in which the cost of veridical strategies is lower (see Payne, Bettman, & Johnson, 1990).[1] Russo (1977), for example, presented supermarket shoppers with organized unit-price lists that were easier to use than item-by-item shelf tags. The change in format engendered a 2% decrease in shoppers' expenditures. In a laboratory task requiring choices among simple gambles, Johnson, Payne, and Bettman (1988) found that changing probability numbers from complicated fractions (e.g., 29/36) to simple decimals (e.g., 0.8) sharply reduced internally inconsistent choices ("preference reversals").

Performance is also affected by how judges encode, store, and retrieve the information they process. It may be possible to restructure the way information is provided so that it better suits the processes applied to it. For example, Klayman (1988) found that subjects were better able to learn from feedback in a multiple-cue task when they were allowed to choose their own combinations of cues to test than when they passively observed a representative set of instances. Another striking example involves hypothesis testing. People have a tendency toward

[1]Payne et al. (1990) use the term "adaptive" to refer to situations in which people profitably modify their strategies in response to local task conditions, such as the amount of time and information available. Thus, their use of the term is not equivalent to Anderson's (1990, 1991), and corresponds more closely to our "adaptable".

"positive testing" (Klayman & Ha, 1987). That is, they test instances they think will fit their hypothesis, paying less attention to instances they think will not fit. Thus, they discover false positives more readily than false negatives, and tend toward overly narrow hypotheses (Klayman & Ha, 1989). Tweney et al. (1980) found a very simple way around this problem. Their subjects were required to infer a rule that designated a subset of instances (certain sets of three numbers, as in Wason, 1960). Instead of classifying instances as those that fit the rule and those that did not, they described the same instances as belonging to one of two mutually exclusive and exhaustive categories, "Med" and "Dax". This resulted in much better performance, probably because, when conceptualized this way, subjects conducted some positive tests of a hypothesized "Med" rule and some of a hypothesized "Dax" rule, and because they then were able to encode each datum as being a positive example of something (either Med or Dax).

Interventions like these do not involve attempts to teach people different ways of reasoning or to give them more information. Instead, the environment is adjusted to fit existing processes. Some interventions (e.g., Russo's manipulation of unit-price displays) work by making it easier to execute a given process well, or facilitating the use of a better one already in the judge's repertoire. Others (e.g., Tweney et al.'s Med–Dax framing) provide an information structure that works better with the processes people use. We do not wish to imply that changing environments instead of processes is always the best approach, or even that it is always feasible. However, in some circumstances it may prove to be the most effective means of improving judgment.

A case study: learning to diagnose

We applied the approach of modifying the environment rather than the cognitive processes to a problem in diagnostic reasoning. Research in a number of different contexts points to a general inferential strategy that can be characterized as follows: given a set of data and one or more hypotheses, people judge the likelihood of a hypothesis by judging how well the observed data conform to the pattern expected under the hypothesis. This inferential strategy has been proposed both as a general heuristic for judgment under uncertainty (an aspect of "representativeness", see Kahneman, Slovic, & Tversky, 1982, Part II) and as a basic mechanism for categorization (see Komatsu, 1992; Osherson, Smith, Wilkie, López, & Shafir, 1990).

It may be that this inferential strategy is generally well adapted to the environment (in the spirit of Anderson, 1991). However, it can also give rise to some systematic errors in diagnostic reasoning (see Fischhoff & Beyth-Marom, 1983). Investigators have expressed concern over related phenomena in medicine (see Elstein, Shulman, & Sprafka, 1978; Schwartz & Griffin, 1986), auditing (see

Smith & Kida, 1991), and law (e.g., Saks & Kidd, 1980). Many of the observed problems stem from the fact that representative features are not necessarily diagnostic, and vice versa. From a Bayesian perspective, a feature is representative or typical if p(Feature|Category) is high, but a feature is diagnostic to the extent that p(Feature|Category) differs from p(Feature|Out-of-category).[2] People sometimes confuse the former with the latter, a problem known as "pseudo-diagnosticity" (Doherty, Mynatt, Tweney, & Schiavo, 1979). This problem reflects a broader tendency among hypothesis testers to focus on the relation of the data to the focal or preferred hypothesis, giving less attention to their relation to alternative hypotheses (Fischhoff & Beyth-Marom, 1983; Klayman & Ha, 1987, 1989; Slowiaczek, Klayman, Sherman, & Skov, 1992).

Suppose, for example, that you are an archeologist who uncovers a piece of pottery that must come from either Shell Island or Coral Island (Doherty et al., 1979; Doherty, Schiavo, Tweney, & Mynatt, 1981). You observe features such as a narrow mouth and curved handles. What information would you most want to help you determine the probability that the pot came from Shell Island? Subjects often choose to ask about the proportion of Shell Island pots with narrow necks and with curved handles, and are less interested in the proportion of those characteristics on Coral Island. In other words, people base their confidence more on the typicality of the feature, p(Feature|Category), than on the likelihood ratio that indicates diagnosticity, p(Feature|Category)/p(Feature|Out-of-category).[3]

Our study uses the context of medical diagnosis (see also Kern & Doherty, 1982; Gruppen, Wolf, & Billi, 1991; Wolf, Gruppen, & Billi, 1985). Consider a fictitious patient known to suffer from one of two fictitious diseases, proxititis and zymosis, described in Table 1. The last feature listed is the type of cough observed. The clinician's concept of proxititis[4] might include the fact that a dry cough was somewhat more common than a harsh cough. Thus, if a harsh cough were observed in a patient, that might be taken as modest evidence against a proxititis diagnosis. From a Bayesian standpoint, that would be a significant error of judgment. Harsh coughs are much more common in proxititis than in zymosis, so that symptom is strong evidence *in favor of* the former disease. Given the

[2]Specifically, Bayes' equation species that $\Omega' = [p(f|c)/p(f|-c)] \cdot \Omega$, where $p(f|c)$ is the probability of observing feature f given that the case belongs to category c, $p(f|-c)$ is the probability of observing f given that the case does not belong to category c, Ω is the odds the judge gives that the case belongs to c before observing f, and Ω' is the odds the judge should give after observing f.

[3]Our discussion ignores two interesting related phenomena. First, we assume that a feature could represent the presence of something or the absence of something. People may not treat these two representations identically (Christensen-Szalanski & Bushyhead, 1983; Sherman & Corty, 1984). Second, Bayesian inference also requires consideration of the base rates of the hypotheses, and people may not always incorporate base rate information properly (see Bar-Hillel, 1990).

[4]We use the term *concept* rather than the more specific *prototype*, because we do not wish to (or need to) take a position in the ongoing controversy over whether category concepts are organized around specific prototypes or generalized representations of exemplars (see Komatsu, 1992; Medin & Smith, 1984; Oden, 1987).

Table 1. *Conditional probabilities of feature given each disease*

Cue	Feature	Proxititis	Zymosis
Age	<18 years	.9	.6
	>18 years	.1	.4
Sex	Male	.3	.7
	Female	.7	.3
Race	White	.6	.9
	Black/brown	.4	.1
Fever	High-grade	.4	.1
	Low-grade	.6	.9
Anorexia	Nausea	.7	.3
	Diarrhea	.3	.7
Prostration	Severe	.1	.4
	Mild	.9	.6
Difficulty walking	Pain	.3	.7
	Tingling	.7	.3
Tongue texture	Smooth	.1	.4
	Cracked	.9	.6
Rash	Moist	.7	.3
	Scaly	.3	.7
Cough	Dry	.6	.9
	Harsh	.4	.1

conditional probabilities shown in Table 1, the odds of proxititis are six times higher with a harsh cough than with a dry one, even though most proxititis patients have dry coughs.[5]

Researchers in medical decision making have found that trained clinicians do, in fact, confuse the typicality of a case feature (or, equivalently, the "sensitivity" of a test) with its diagnosticity. This is especially so of novices (Bordage & Lemieux, 1991; Feltovich, Johnson, Moller, & Swanson, 1984; Patel, Evans, & Kaufman, 1989; Lemieux & Bordage, 1992). This tendency toward pseudodiagnostic reasoning may be fostered, or at least reinforced, by the way in which medical knowledge is presented. Medical textbooks are usually organized around particular diseases or categories of diseases, presenting the symptomatology of each and the pathophysiological principles that explain the symptoms (Bordage & Lemieux, 1990; Kriel & A'Beckett-Hewson, 1986). Similarly, in ward rounds and grand rounds, much of the instruction revolves around a "disease of the week", with discussion of symptomatology and presentations of illustrative cases.

The upshot is that medical students are generally trained to know the pattern of signs and symptoms associated with a particular disease. However, the diagnostic task is to find the disease associated with a particular set of signs and

[5]The odds of proxititis (abbreviated prox) given a harsh cough are equal to the prior odds of prox (Ω_{prox}) times the likelihood ratio, $p(\text{harsh}|\text{prox})/p(\text{harsh}|\text{zymosis})$. This equals $(.4/.1) \cdot \Omega_{\text{prox}} = 4 \cdot \Omega_{\text{prox}}$. The odds of prox given a dry cough are $\Omega_{\text{prox}} \cdot p(\text{dry}|\text{prox})/p(\text{dry}|\text{zymosis}) = (.6/.9) \cdot \Omega_{\text{prox}} = 2/3 \cdot \Omega_{\text{prox}}$.

symptoms (i.e., the presenting problems of the patient). As Barrows and Bennett (1972) put it, the student must "synthesize all this information and apply it in reverse when confronted with a ... patient [who] never presents as a textbook case of anything" (p. 277). Or, as Kriel and A'Beckett-Hewson (1986) state, "medical textbooks and classroom teaching abound in the presentation of detailed lists of disorders, features and therapeutic actions which fail to provide a categorization scheme that is best suited for their retrieval in a clinical problem-solving situation" (p. 95). The complaint is not that students receive too little information, or the wrong information. Rather, the worry is that medical education does not help students to process information in a manner that suits the task of diagnosis.

There have been a number of proposals to change medical training to improve diagnostic performance. The goal is generally to improve clinicians' reasoning processes. One major reform effort centers on the use of "problem-based learning" (Albanese & Mitchell, 1993; Barrows, 1986; Barrows & Tamblyn, 1980; Berkson, in press; Schmidt, Dauphinee, & Patel, 1987). Problem-based curricula focus on real or realistic cases. The instructor works through the process of diagnosis with students, using the clinical problems as a vehicle for integrating basic science and clinical practice. A problem-based approach provides opportunities to train students to avoid pseudodiagnostic thinking. For example, teachers could prompt students to cross-reference different diseases by asking questions such as, "What would you say if the patient had symptoms X and Y instead of A and B?" or "How can you tell this is rheumatoid arthritis, rather than osteoarthritis?" Generally, though, such pedagogic techniques are not a systematic part of the curriculum, and it is not clear how widely they are applied. Furthermore, problem-based curricula tend to present a long series of cases of one particular condition, instead of a mixed series including cases of other kinds (Norman, 1988). Overall, there is little evidence that problem-based learning effectively improves diagnostic accuracy (Albanese & Mitchell, 1993; Berkson, in press; Norman & Schmidt, 1992; Patel, Groen, & Norman, 1991; Schmidt, Dauphinee, & Patel, 1987).

Other educators recommend training students in the use of systematic methods of hypothetico-deductive reasoning (see Elstein, Shulman, & Sprafka, 1990; Kassirer, 1983). There is little evidence for or against the effectiveness of such direct training in the medical domain. However, several studies suggest that use of proper hypothetico-deductive logic does not differ appreciably between expert and non-expert diagnosticians, nor between more- and less-successful diagnosticians (Groen & Patel, 1985; Neufeld, Norman, Barrows, & Feightner, 1981; Norman, Tugwell, Feightner, Muzzin, & Jacoby, 1985). This suggests that training in logical techniques may not be providing a critical missing skill. Others have advocated training clinicians in the use of formal, normative techniques such as Bayes' theorem and decision analysis (Schwartz, Gorry, Kassirer, & Essig,

1973; Sox, Blatt, Higgins, & Marton, 1988; Weinstein & Fineberg, 1980). Here, too, there is little evidence about the effects of such training on practice.

In medical education, as in other domains, recommendations for improving judgment skills focus on changing the reasoning processes to better fit the task. In our study, we took the opposite tack, taking for granted the reasoning process that engenders pseudodiagnosticity, and redesigning the information provided to reduce the errors produced by those processes. We assumed that clinicians judge the likelihood of a diagnosis by testing how well the observed case features match the features associated with their concept of that disease. Judgments based on such a heuristic process could be much more accurate if the clinicians' disease concepts included lists of *distinctive* features, rather than lists of *typical* features. For example, the concept of proxititis would indicate that harsh coughs were much more common in that disease than in zymosis, and dry coughs somewhat less common than in zymosis. Then, harsh coughs, rather than dry coughs, would be associated with proxititis, and a patient's harsh cough would be seen, properly, as a good match to proxititis. Deviations from proper Bayesian thinking might still remain (e.g., misestimations of the relative strengths of different items of evidence, or insufficient attention to disease base rates). However, diagnostic judgments should be much closer to the Bayesian ideal than if one is *taught* what is typical, and must *infer* what is distinctive.

One way to accomplish this transformation might be to change the structure of information with which student diagnosticians have to work. Instead of disease concepts based on the likelihoods of features given a disease, students could learn concepts based on features that distinguish one disease from another. Cognitive research on categorization suggests that these sorts of "contrastive" concepts are not difficult to learn. The category concepts people develop from everyday experience are based on a combination of typicality and discriminability (Rosch & Mervis, 1975), and experienced clinicians develop disease concepts that include discriminant features (see Johnson, 1982).

To encourage the formation of contrastive concepts, training should emphasize the ways in which the disease in question differs from other diseases that are frequently alternative explanations for similar patterns of findings. An objection might be raised that learning contrastive concepts would be impractical because, for any n diseases, it would be necessary to learn $n(n-1)$ sets of contrasts between diseases. Two factors mitigate this problem. First, diseases tend to fall into relatively small sets of logical competitors (two, three, or four), so in practice, n is not usually large. Second, contrasts need not be pairwise. Knowledge of diseases can be effectively organized into hierarchical groupings of related diseases (Beck & Bergman, 1986; Feltovitch et al., 1984; Lemieux & Bordage, 1992; Politser, 1987; Wortman, 1972); then each particular disease need be contrasted only with the average of the group (e.g., "harsh cough is seen more often in proxititis than is usual for this class of diseases"). Contrastive concepts

would be especially useful if they included direct information about the degree of diagnosticity of each feature (i.e., likelihood ratios such as "four times as likely in proxititis as in . . ."); however, the formation of contrastive concepts does not require any information that is not required for independent concepts. The key difference is one of structure rather than content.

In the present study, we demonstrate the potential benefits of a change in information structure by teaching subjects to recognize cases of the two fictitious diseases referred to in Table 1: proxititis and zymosis. Half the subjects learned each disease separately, using written materials and examples similar to those used in traditional medical education. The other half of the subjects received exactly the same information, but the two diseases were presented in parallel, so that subjects could easily infer the differences between them.

We hypothesized that training the two diseases independently would encourage the formation of independent concepts based on those features that are most frequent in patients with the given diseases. Juxtaposing the two diseases during training should encourage attention to *relative* typicality, and should thus produce concepts that incorporate contrastive information. We did not expect contrastive training to change the basic method by which subjects evaluated hypotheses. In judging how typical a given case is of one disease, or in making a choice between alternative diagnoses, we expected that subjects in both conditions would base their judgments on how well the features of the case matched the features associated with the disease concept. However, we expected subjects in the two conditions to form different disease concepts. Thus, both typicality judgments and differential diagnoses should show differences between the two training methods because subjects are matching to different concepts. Moreover, because contrastive concepts are closer to the Bayesian sense of diagnosticity based on likelihood ratios, the diagnostic judgments of the contrastive training group should be significantly closer to the statistically correct diagnostic judgments.

Method

Subjects

Subjects were 48 students at the University of Chicago who responded to advertisements placed on campus noticeboards. Subjects received a fixed fee for participating in the experiment, which took about 45 min to complete. The procedure was administered in groups of two to five subjects, with each group assigned randomly to either the independent training condition or the contrastive training condition (24 subjects in each). Within each training condition, 12 subjects performed a typicality judgment task, and 12 performed a categorization task.

Materials

Subjects were provided with an introductory information sheet indicating that the subject's goal was to learn to recognize cases of two equally common, fictitious diseases: proxititis and zymosis. They were also given a summary sheet listing the ten cues about which information would be available, and the two possible values that each cue might take (resembling Table 1, without the probabilities). The training procedure included written information about the diseases (described later), and 30 slides of cases of each disease (an example is provided in Fig. 1). The set of profiles for each disease was designed to reflect closely the conditional probabilities of each feature given the disease.

An additional set of 20 profiles was printed on separate sheets of paper for use in the judgment tasks following training. These cases were designed to match independent and contrastive disease concepts in different ways. In modeling the degree of match between a case and a concept, we assumed that the independent concept for a given disease comprised those features that were present more often than not for that disease, and the contrastive concept comprised those features that were more common in that disease than in the other. We then counted the number of case features that matched the features in each concept, ignoring the possibility of differential weighting of different features. Consider the case described in Table 2. According to our simple feature-matching model, this case fits the independent concept of proxititis better than zymosis (designate this $\text{Prox}_{\text{indep}}$). From the point of view of contrastive concepts, however, the fit is equal to each disease (designated $\text{Equal}_{\text{contr}}$). Thus, the case described in Table 2 is classified as $\text{Prox}_{\text{indep}}/\text{Equal}_{\text{contr}}$.

Because contrastive concepts are based on features that are diagnostic of the

<u>ZYMOSIS</u>

Age:	11 YEARS
Sex:	MALE
Race:	WHITE
Fever:	LOW-GRADE
Reason for anorexia:	DIARRHEA
Degree of prostration:	MILD
Difficulty walking due to:	TINGLING SENSATION
Texture of tongue:	CRACKED AND FISSURED
Appearance of rash:	SCALY
Type of cough:	HARSH, BARKING

Figure 1. *One of the sample cases presented to subjects during training.*

Table 2. *Example of a case for which judgments based on independent concepts and contrastive concepts differ*

		Matches to independent concepts		Matches to contrastive concepts	
		Prox.	Zym.	Prox.	Zym.
Age	>18 years	−	−	−	+
Sex	Female	+	−	+	−
Race	White	+	+	−	+
Fever	Low-grade	+	+	−	+
Anorexia	Nausea	+	−	+	−
Prostration	Mild	+	+	+	−
Walking	Tingling	+	−	+	−
Tongue	Smooth	−	−	−	+
Rash	Moist	+	−	+	−
Cough	Dry	+	+	−	+
Matches to concept		8	4	5	5

disease, rather than features that are common given the disease, matching to contrastive concepts produces a fair approximation to Bayesian diagnostic judgment. In all of the presented cases, the disease favored by matching to contrastive concepts was also the more likely diagnosis by Bayesian calculation, and when the fit to the two disease concepts was equal, the probabilities of the two diseases were indeed close. Seven types of profiles were constructed, as summarized in Table 3.

Table 3. *Types of cases used in typicality judgment and differential diagnosis tasks*

Type of cases[a]	Number of cases	Match to independent concepts	Match to contrastive concepts
$\text{Prox}_{\text{indep}}/\text{Equal}_{\text{contr}}$	4	proxititis > zymosis	proxititis = zymosis
$\text{Zym}_{\text{indep}}/\text{Equal}_{\text{contr}}$	4	zymosis > proxititis	zymosis = proxititis
$\text{Equal}_{\text{indep}}/\text{Prox}_{\text{contr}}$	4	proxititis = zymosis	proxititis > zymosis
$\text{Equal}_{\text{indep}}/\text{Zym}_{\text{contr}}$	4	zymosis = proxititis	zymosis > proxititis
$\text{Equal}_{\text{indep}}/\text{Equal}_{\text{contr}}$	2	proxititis = zymosis	proxititis = zymosis
$\text{Prox}_{\text{indep}}/\text{Prox}_{\text{contr}}$	1	proxititis > zymosis	proxititis > zymosis
$\text{Zym}_{\text{indep}}/\text{Zym}_{\text{contr}}$	1	zymosis > proxititis	zymosis > proxititis

[a]In labeling the types of cases, Prox(ititis) and Zym(osis) indicate which disease concept better fits the features of the case, according to either indep(endent) or contr(astive) concepts. "Equal" indicates that the case features are about equally well matched to the proxititis and zymosis concepts of that type.
Note: In all cases, ">" indicates that the match to the left-hand disease was better by four features (e.g., two features in the profile matched the proxititis concept, six matched zymosis), and "=" indicates an equal number of feature matches to each concept.

Training procedures

Independent training condition

Subjects in the independent training condition began by reading a "summary chapter" that described one of two fictitious diseases, either "proxititis" or "zymosis". They were instructed to study this summary chapter carefully and to form a general impression of what a case of this disease looked like, based on the general description and some case histories which they would be shown.

Each summary chapter provided information about the signs and symptoms associated with the disease (see Appendix A). These chapters were similar to descriptions found in medical digests. In order to parallel the type of information provided in medical texts, probabilities and frequencies were expressed verbally rather than numerically. The translation from the numerical probabilities shown in Table 1 to verbal expressions was done in accord with studies of the meaning of probability-related words (Beyth-Marom, 1982; Kenney, 1981; Nakao & Axelrod, 1983; Wallsten & Budescu, 1983). Additionally, each of the diseases was described as being of "relatively frequent" occurrence, so no differences in base rates were indicated. Neither of the introductory chapters made any reference to the other disease.

Subjects were given 5 min to study the summary chapter. Then they were shown a series of 30 slides presenting case histories of the disease they had just read about, as in Fig. 1. Prior to viewing the first slide subjects were reminded that their objective was to be able to recognize a case of the disease when they saw one and that they should form a general impression of patients with the disease rather than try to remember individual patients. They viewed the slides in random order for 10 s each, without taking notes. Then subjects were given a summary chapter and 30 case histories for the second disease, following the same procedure. Half the subjects were trained on proxititis first, half on zymosis. After this second set of presentations, subjects were asked to list the features of a typical case of the first disease, and then, separately, the features of a typical case of the second disease.

After subjects had finished writing descriptions of a typical case of proxititis and a typical case of zymosis, half were given a typicality judgment task and half were given a classification task. Subjects completed these tasks at their own pace.

Contrastive training condition

The information provided in contrastive training was the same as in independent training. The two conditions differed in how the information was organized in the written summaries and in the case presentations. Subjects in the contrastive condition began by reading a summary chapter that described the signs and symptoms associated with both diseases. In essence, this chapter was an

amalgam of the two chapters seen by subjects in the independent condition (see Appendix B). Contrastive subjects were given 10 min to study the chapter, with instructions to form a general impression of what a case of proxititis looked like and what a case of zymosis looked like, on the basis of this summary chapter and the set of case histories they would see. They saw the same 60 case histories that their counterparts in the independent condition saw, except that proxititis and zymosis patients were randomly intermixed during a single viewing session. Otherwise the procedure followed was the same as that for subjects in the other training condition.

Judgment task procedures

Twelve subjects in each training condition rated 10 patient profiles on a 7-point scale with endpoints labeled "not at all like a typical case of [disease]" and "very much like a typical case of [disease]". They were given printed case summaries, similar in format to the slides they had viewed (see Fig. 1). Each case appeared on a separate page, with the scale printed at the bottom. The subjects first rated the 10 cases with regard to one disease, and then the cases were presented again in a different random order to be rated with respect to the other disease. The 10 cases were a subset of the 20 listed in Table 3: two each of types $\text{Prox}_{\text{indep}}/\text{Equal}_{\text{contr}}$, $\text{Zym}_{\text{indep}}/\text{Equal}_{\text{contr}}$, $\text{Equal}_{\text{indep}}/\text{Prox}_{\text{contr}}$, $\text{Equal}_{\text{indep}}/\text{Zym}_{\text{contr}}$, and $\text{Equal}_{\text{indep}}/\text{Equal}_{\text{contr}}$. The other 12 subjects in each training condition made different differential diagnoses. They received case summaries with the two disease names printed at the bottom instead of the rating scale. (The order of the names was counterbalanced across subjects.) Subjects indicated which of the two diseases they believed to be the correct diagnosis for each of the 20 cases listed in Table 3.

Results

Disease descriptions

The first hypothesis to be tested is that the structure of information presentation in training affects the nature of the category concepts formed by learners. We examined the lists of features subjects provided when asked to describe a typical case of each disease. In the independent condition, we expected that subjects' descriptions would be based on absolute typicality. In other words, the description of proxititis would consist of those features found in the majority of proxititis patients. In the contrastive condition, however, we expected that subjects would be influenced by *relative* typicality, and the description of proxititis would be more likely to include features that were more frequent in proxititis than in the

alternative, zymosis. The feature discussed earlier, type of cough, is an example: proxititis patients are more likely to have a dry cough than a harsh one, but having a harsh cough is more common with proxititis than with zymosis (see Table 1). Thus, independent subjects should describe the typical proxititis patient as having a dry cough, but contrastive subjects may instead identify proxititis with a harsh cough.

Table 4 shows which features subjects included in their descriptions of the typical case of proxititis and zymosis. For the first three cues (race, fever, and cough), absolute and relative typicality pointed to opposite features in proxititis; for the next three (age, prostration, and tongue condition), absolute and relative typicality differed in zymosis. These six cues provided data for three planned comparisons. In the independent condition, subjects were significantly more likely to report features consistent with absolute typicality than with relative typicality, $t(23) = 2.24$, $p < .05$. On average, they named the feature with higher absolute typicality 3.5 times out of 6, and the feature with higher relative typicality 2.2

Table 4. *Number of subjects who included each value in descriptions of the typical case of each disease, by training condition*

		Proxititis		Zymosis	
Cue	Feature	Indep.	Contrast.	Indep.	Contrast.
Race[a]	White	14	7	15	16
	Black/brown	8	17	7	5
Fever[a]	Low	13	7	16	16
	High	9	15	6	5
Cough[a]	Dry	14	7	18	18
	Harsh	10	13	5	6
Age[b]	<18	17	17	14	7
	>18	7	2	8	15
Prostration[b]	Mild	19	14	15	9
	Severe	4	9	9	13
Tongue[b]	Smooth	9	9	10	13
	Cracked	13	13	13	10
Sex[c]	Female	19	20	8	3
	Male	5	4	15	20
Anorexia[c]	Diarrhea	7	8	16	19
	Nausea	15	15	7	5
Walking[c]	Pain	4	3	19	18
	Tingling	20	19	4	6
Rash[c]	Scaly	6	6	17	12
	Moist	17	17	7	11

Note: Numbers exclude instances in which subjects mentioned both features or neither.
[a]Contrastive and independent conditions predicted to differ for proxititis.
[b]Contrastive and independent conditions predicted to differ for zymosis.
[c]Contrastive and independent conditions predicted to agree within each disease.

times. (The other 0.3 times the descriptions included both features or neither.) In the contrastive condition, subjects named the absolute-typicality feature an average of 1.9 times, and the relative-typicality features 3.6 times, $t(23) = 4.79$, $p < .001$. The interaction between condition and feature choice was significant, $t(46) = 4.40$, $p < .001$.

These results indicate that the independent and contrastive modes of presentation produced qualitatively different independent and contrastive disease concepts. In the independent condition, subjects most often identified a disease with those features that were present in a majority of cases. Contrastive-condition subjects were strongly influenced by the relative frequency of occurrence of a feature in one disease versus the other: they were more likely to identify a disease with the minority feature than the majority feature if the minority feature had a higher relative frequency.

Judgments of typicality

Ratings of the typicality of individual cases provide further tests of the hypothesis that independent and contrastive training produce qualitatively different disease concepts. We assume that typicality ratings are influenced primarily by the extent to which features in the patient profile match the features of the disease concept. Thus, differences in disease concepts should produce differences in the typicality ratings provided by subjects in the two training conditions.

Recall that subjects rated five kinds of cases ($\text{Prox}_{\text{indep}}/\text{Equal}_{\text{contr}}$, $\text{Zym}_{\text{indep}}/\text{Equal}_{\text{contr}}$, $\text{Equal}_{\text{indep}}/\text{Prox}_{\text{contr}}$, $\text{Equal}_{\text{indep}}/\text{Zym}_{\text{contr}}$, and $\text{Equal}_{\text{indep}}/\text{Equal}_{\text{contr}}$) with respect to each disease. If subjects in the independent condition are matching to independent concepts, they should judge $\text{Prox}_{\text{indep}}/\text{Equal}_{\text{contr}}$ cases to be more typical of proxititis than of zymosis, and vice versa for $\text{Zym}_{\text{indep}}/\text{Equal}_{\text{contr}}$ cases. For the other three kinds of cases, they should rate the two diseases about equally. As shown in Table 5, the results fit the predictions for all five types of cases. If subjects in the contrastive condition are matching to contrastive concepts, they should rate proxititis higher in $\text{Equal}_{\text{indep}}/\text{Prox}_{\text{contr}}$ cases and zymosis higher in $\text{Equal}_{\text{indep}}/\text{Zym}_{\text{contr}}$ cases, and there should be no difference in the other three types of cases. In the contrastive condition, the predicted pattern was observed in four of five types of cases, the exception being the significant difference in $\text{Prox}_{\text{indep}}/\text{Equal}_{\text{contr}}$ cases.

These findings suggest that subjects in each condition are matching to different concepts. It is also possible to test condition differences directly, by testing for Disease × Condition interactions. Such interactions should exist for all types of cases except $\text{Equal}_{\text{indep}}/\text{Equal}_{\text{contr}}$, with subjects in one condition rating the two diseases about equal, and subjects in the other condition rating one disease higher than the other. The predicted interaction was significant in $\text{Zym}_{\text{indep}}/\text{Equal}_{\text{contr}}$

Table 5. *Ratings of typicality of cases with respect to proxititis and zymosis, by training condition and type of case*

Training condition	Disease rated	Type of case: $Prox_{indep}/Equal_{contr}$	$Zym_{indep}/Equal_{contr}$	$Equal_{indep}/Prox_{contr}$	$Equal_{indep}/Zym_{contr}$	$Equal_{indep}/Equal_{contr}$
Independent	Proxititis	5.1	3.0	4.4	4.2	3.8
	Zymosis	2.7**	4.8**	3.6[a]	3.9[a]	3.7[a]
Contrastive	Proxititis	4.6	4.2	5.1	3.7	4.0
	Zymosis	3.5*	4.3[a]	3.6**	5.0**	4.1[a]

*Proxititis and zymosis ratings differ, $p < .05$.
**Ratings differ, $p < .005$.
[a]Ratings not significantly different, $p > .10$.

cases, $F(1, 22) = 8.72$, $p < .008$, and $Equal_{indep}/Zym_{contr}$ cases, $F(1, 22) = 12.64$, $p < .002$, marginally significant in $Prox_{indep}/Equal_{contr}$ cases, $F(1, 22) = 4.21$, $p < .06$, and nonsignificant in $Equal_{indep}/Prox_{contr}$ cases, $F(1, 22) = 1.35$, $p < .26$.

In sum, despite some evidence of overlap between the concepts formed in independent and contrastive groups, the results provide strong support for the conclusion that independent and contrastive training produced different disease concepts.

Differential diagnosis

Subjects in the independent and contrastive training conditions developed different concepts of what is typical for each disease, but how does this affect diagnostic judgment? Studies of representativeness and pseudodiagnosticity suggest that subjects will choose the disease with the best match between typical features and the features observed in the case. If so, then the different concepts formed by subjects in the independent and contrastive conditions should produce different diagnostic judgments. Data to test this hypothesis come from the differential diagnoses of 20 cases made by 12 subjects in each condition. These analyses use the 16 cases for which condition differences are predicted: four each of types $Zym_{indep}/Equal_{contr}$, $Prox_{indep}/Equal_{contr}$, $Equal_{indep}/Zym_{contr}$, and $Equal_{indep}/Prox_{contr}$.

As with typicality ratings, our feature-matching hypothesis predicts that subjects will find that some cases fit one disease much better than the other, whereas other cases fit both diseases about equally. When one disease matches better than the other there should be good consensus as to which are proxititis cases and which are zymosis; when they match about equally there should be no clear distinction. As before, whether a case is clear or equivocal should depend on which kind of disease concepts are used.

We analyzed the average diagnoses proportion of cases in which proxititis was the diagnosis. (The proportion of zymosis diagnoses was one minus this; there were no missing judgments.) In accord with predictions, independent subjects diagnosed proxititis more often in $\text{Prox}_{\text{indep}}/\text{Equal}_{\text{contr}}$ cases ($M = .72$) than in $\text{Zym}_{\text{indep}}/\text{Equal}_{\text{contr}}$ cases ($M = .42$), $t(11) = 3.19$, $p < .009$, but contrastive subjects did not distinguish the two types ($M = .54$ and .58, respectively), $t < 1$. The interaction between condition and case type was significant in a 2×2 ANOVA, $F(1, 22) = 8.59$, $p < .008$. Similarly, contrastive subjects chose proxititis more often in $\text{Equal}_{\text{indep}}/\text{Prox}_{\text{contr}}$ cases ($M = .71$) than in $\text{Equal}_{\text{indep}}/\text{Zym}_{\text{contr}}$ cases ($M = .38$), $t(11) = 3.75$, $p < .004$, but independent subjects did not ($M = .48$ and .52, respectively), $t < 1$. The Case-type $\times$ Condition interaction was again significant, $F(1, 22) = 10.48$, $p < .004$.

We also compared responses to each of the four case types in each condition to chance (.50). These are more stringent tests of our predictions. Of the eight comparisons, there were four in which one diagnosis was expected to predominate. Subjects made the predicted diagnosis more often than chance ($p < .05$) in three of these four instances: $\text{Prox}_{\text{indep}}/\text{Equal}_{\text{contr}}$ in the independent condition, and $\text{Equal}_{\text{indep}}/\text{Prox}_{\text{contr}}$ and $\text{Equal}_{\text{indep}}/\text{Zym}_{\text{contr}}$ in the contrastive condition. The exception was that the independent group found $\text{Zym}_{\text{indep}}/\text{Equal}_{\text{contr}}$ cases to be somewhat equivocal. The proportion of proxititis and zymosis choices did not differ significantly in any of the four instances that were expected to be equivocal.

The data on differential diagnostic judgments support our three basic hypotheses: (a) subjects make differential diagnoses by judging relative goodness of fit between features of the case and features of each disease concept; (b) the structure of information during training affects the concepts applied in diagnosis; and (c) contrastive concepts produce diagnoses that are closer to the statistically appropriate judgments.

Discussion

Our study provides some good news and some bad news about diagnostic reasoning. First the bad news. Novice subjects are not good at translating knowledge about what is typical into knowledge about what is diagnostic. Our results accord with a basic process that has been documented in a variety of reasoning tasks: People judge the likelihood that an instance is a member of a category by how well the features of the instance match the features seen as average or typical for the category.

The good news is that the negative consequences of this feature-matching process may be alleviated by a fairly straightforward change in the way information is presented during instruction. Feature matching is not such a bad method of

making diagnostic judgments if one matches to diagnostic features rather than typical features.

It is not easy to change the basic reasoning processes by which people make diagnostic judgments. The results of the present study suggest that it may not be so hard to change the form of the concepts upon which those processes operate. A change from independent to contrastive concepts was accomplished with a simple restructuring of information, presenting information about two diseases in parallel within one descriptive text and one representative sample of cases. No information was used beyond that provided in independent concepts, taught one disease at a time. Nevertheless, subjects who received the contrastive training were more likely to associate each disease with its discriminant features than with its typical features, and their diagnostic judgments were significantly closer to normative standards.

The task we used is of course a very simple one compared to the complexities of training in medicine or other domains of expert judgment. We cannot be certain that our technique will scale up easily from two diseases and ten features learned in a half hour to hundreds of diseases and features learned in 2 years of preclinical training, or to similarly complex knowledge in other domains. As mentioned earlier, a larger contrastive structure would require a hierarchical organization to avoid the need for innumerable pairwise contrasts. However, our results suggest a direction worth trying. In medical education, for example, teaching basic disease information in a contrastive format could provide novices with a knowledge structure that is closer to that of experts and better adapted to the task of differential diagnosis. A contrastive knowledge structure might also put students in a better position to profit from problem-based training and instruction in Bayesian principles, because there would be less discrepancy between the concepts being taught and the structure of information in memory.

Conclusions

Non-optimalities in judgment can be thought of as mismatches between people's environments and the processes they use to attain their goals in those environments. A better match between processes and environments can be achieved through changes in either or both. Most efforts to improve judgment focus on changing cognitive processes or enhancing them through the use of analytical tools. There are a number of potential advantages to that approach. Under favorable circumstances, training to avoid errors in one context may generalize to other contexts. For example, a person who learned to think about diagnosticity rather than typicality in diagnosing diseases might carry the lesson to problems in automobile repair or human relations. Furthermore, training in more effective processes could yield more wide-ranging benefits. For example, a skilled

Bayesian reasoner would presumably avoid the ill effects of pseudodiagnosticity, base rate neglect, and conjunction fallacies simultaneously.

On the other hand, it may sometimes be more effective to take the processes as given, and think about how the environment can be adapted to suit them. We advocate devoting more attention to this approach as an alternative means of intervention. Given the non-intuitive nature of many statistical principles and decision analysis techniques, it may often be easier to modify the environment than the person. Moreover, whereas better thinking must be taught to each individual, the benefits of a modified environment are likely to generalize across the range of individuals who encounter it.

To be effective, though, improvements in the environment must be targeted to specific goals. In our study, for example, teaching contrastive concepts without changing the basic judgment process might reduce pseudodiagnosticity, but create "pseudotypicality". That is, contrastive-trained subjects may identify a low-frequency feature as typical of a disease if its frequency given the other diseases is lower. Where accurate diagnostic judgment is a more central goal than accurate judgments of typicality, the trade-off between pseudodiagnosticity and pseudotypicality would be favorable.

Thinking about the design of favorable environments highlights the importance of research into environment–person interactions. When and how do people adapt their cognitive processes in response to environmental characteristics? The answers to this question can guide the design of environments that facilitate the use of better strategies. How do environmental characteristics affect the performance of different processes? Answers here can guide the design of the environmental inputs to human cognition to obtain the best performance from given strategies.

Some critics of "heuristics and biases research" complain that researchers are too interested in demonstrating irrationality, and are not impressed enough with what well-fashioned cognitive machines people are. Indeed, many of the biases, fallacies, and imperfections of human judgment and reasoning seem to come and go, depending on the wording of the problem, the goals of the subject, the subject's expertise in the problem domain, etc. One possible interpretation is that the purported biases are not robust, or are artifacts of the laboratory. An alternative view is this: most of the environments people face in life are artificial; a great many are ambiguously defined, present conflicting goals, and are unfamiliar. Sometimes people manage quite well, by standards that are important to them, sometimes they do not.

It is reasonable to expect human cognition to be well adapted to the general environment, but not completely adaptable to each subenvironment encountered. One can recognize human judgmental errors as real and perhaps remediable, while also acknowledging that evolution, social transmission, and learning are powerful forces for shaping appropriate behavior. In other words, it is possible to

be very impressed with both the successes and the failures of human judgment, and to wonder why there is not more of the former and less of the latter. Knowing more about what makes the difference is important for being able to design effective interventions that can improve performance. And seeing what does and does not aid performance can reveal a good deal about underlying processes, too.

References

Abelson, R.P., & Levi, A (1985). Decision making and decision theory. In G. Lindzey & E. Aronson (Eds.), *The Handbook of Social Psychology* (3rd ed., Vol. 1, pp. 231–309). New York: Random House.

Albanese, M.A., & Mitchell, S. (1993). Problem-based learning: a review of literature on its outcomes and implementation issues. *Academic Medicine*, *68*, 52–81.

Anderson, J.R. (1990). *The Adaptive Character of Thought*. Hillsdale, NJ: Erlbaum.

Anderson, J.R. (1991). Is human cognition adaptive? *Behavioral and Brain Sciences*, *14*, 471–517.

Bar-Hillel, M. (1990). Back to base rates. In R. Hogarth (Ed.), *Insights in decision making: a tribute to Hillel J. Einhorn* (pp. 200–216). Chicago: University of Chicago Press.

Barrows, H.S. (1986). A taxonomy of problem-based learning methods. *Medical Education*, *20*, 481–486.

Barrows, H.S., & Bennett, K. (1972). The diagnostic (problem solving) skill of the neurologist. *Archives of Neurology*, *26*, 273–277.

Barrows, H.S., & Tamblyn, R.N. (1980). *Problem-based learning: an approach to medical education*. New York: Springer.

Beck, A.L., & Bergman, D.A. (1986). Using structured medical information to improve students' problem-solving performance. *Journal of Medical Education*, *61*, 749–756.

Berkson, L. (in press). Problem-based learning: have the expectations been met? *Academic Medicine*.

Beyth-Marom, R. (1982). How probable is probable? A numerical translation of verbal probability expressions. *Journal of Forecasting*, *1*, 257–269.

Beyth-Marom, R., Dekel, S., Gombo, R., & Shaked, M. (1985). *An elementary approach to thinking under uncertainty*. Hillsdale, NJ: Erlbaum.

Bordage, G., & Lemieux, M. (1990). Which medical textbook to read? Emphasizing semantic structures. *Academic Medicine*, *65*, S23–S24.

Bordage, G., & Lemieux, M. (1991). Semantic structures and diagnostic thinking of experts and novices. *Academic Medicine*, *66*, S70–S72.

Christensen-Szalanski, J.J.J., & Bushyhead, J.B. (1983). Physicians' misunderstanding of normal findings. *Medical Decision Making*, *3*, 169–175.

Cohen, L.J. (1981). Can human irrationality be experimentally demonstrated? *Behavioral and Brain Sciences*, *4*, 317–331.

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

Doherty, M.E., Schiavo, M.D., Tweney, R.D., & Mynatt, C.R. (1981). The influence of feedback and diagnostic data on pseudodiagnosticity. *Bulletin of the Psychonomic Society*, *18*, 191–194.

Dunbar, K. (in press). How scientists really reason: scientific reasoning in real-world laboratories. In R.J. Sternberg & J. Davidson (Eds.), *Insight*. Cambridge, MA: MIT Press.

Edwards, W. (1990). Unfinished tasks: a research agenda for behavioral decision theory. In R. Hogarth (Ed.), *Insights in decision making: a tribute to Hillel J. Einhorn* (pp. 44–65). Chicago: University of Chicago Press.

Elstein, A.S., Shulman, L.S., & Sprafka, S.A. (1978). *Medical problem solving: an analysis of clinical reasoning*. Cambridge, MA: Harvard University Press.

Elstein, A.S., Shulman, L.S., & Sprafka, S.A. (1990). Medical problem solving: a ten year retrospective. *Evaluation and the Health Professions*, *13*, 5–36.

Feltovitch, P.J., Johnson, P.E., Moller, J.H. & Swanson, D.B. (1984). The role and development of medical knowledge in diagnostic expertise. In W. Clancey & E.H. Shortliffe (Eds.), *Readings in medical artificial intelligence* (pp. 275–319). New York: Addison-Wesley.
Fischhoff, B. (1982). Debiasing. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: heuristics and biases* (pp. 422–444). New York: Cambridge University Press.
Fischhoff, B., & Beyth-Marom, R. (1983). Hypothesis evaluation from a Bayesian perspective. *Psychological Review*, *90*, 239–260.
Fong, G.T., Krantz, D.H., & Nisbett, R.E. (1986). The effects of statistical training on thinking about everyday problems. *Cognitive Psychology*, *18*, 253–292.
Funder, D.C. (1987). Errors and mistakes: evaluating the accuracy of social judgment. *Psychological Bulletin*, *101*, 75–90.
Gaeth, G.J., & Shanteau, J. (1984). Reducing the influence of irrelevant information on experienced decision makers. *Organizational Behavior and Human Performance*, *33*, 263–282.
Groen, G.J., & Patel, V.L. (1985). Medical problem-solving: some questionable assumptions. *Medical Education*, *19*, 95–100.
Gruppen, L.D., Wolf, F.M., & Billi, J.E. (1991). Information gathering and integration as sources of error in diagnostic decision making. *Medical Decision Making*, *11*, 233–239.
Hammond, K.R. (1990). Functionalism and illusionism: can integration be usefully achieved? In R. Hogarth (Ed.), *Insights in decision making: a tribute to Hillel J. Einhorn* (pp. 227–261). Chicago: University of Chicago Press.
Hammond, K.R., Hamm, R.M., Grassia, J., & Pearson, T. (1987). Direct comparison of the efficacy of intuitive and analytical cognition in expert judgment. *IEEE Transactions on Systems, Man, and Cybernetics*, *SMC-17*, 753–770.
Hogarth, R.M. (1981). Beyond discrete biases: functional and dysfunctional aspects of judgmental heuristics. *Psychological Bulletin*, *90*, 197–217.
Johnson, E.J., Payne, J.W., & Bettman, J.R. (1988). Information displays and preference reversals. *Organizational Behavior and Human Decision Processes*, *42*, 1–21.
Johnson, P.E. (1982). Cognitive models of medical problem solvers. In D.C. Connelly, E. Benson, & D. Burke (Eds.), *Clinical decision making and laboratory use* (pp. 39–52). Minneapolis: University of Minnesota Press.
Jungerman, H. (1983). The two camps on rationality. In R.W. Scholz (Ed.), *Decision making under uncertainty* (pp. 63–86). Amsterdam: Elsevier.
Kahneman, D., Slovic, P., & Tversky, A. (Eds.) (1982). *Judgment under uncertainty: heuristics and biases*. New York: Cambridge University Press.
Kahneman, D., & Tversky, A. (1979). Intuitive predictions: biases and corrective procedures. *TIMS Studies in Management Science*, *12*, 313–327.
Kassirer, J.P. (1983). Teaching clinical medicine by iterative hypothesis testing: let's practice what we preach. *New England Journal of Medicine*, *309*, 921–923.
Kenney, R.M. (1981). Between never and always. *New England Journal of Medicine*, *305*, 1097–1098.
Kern, L., & Doherty, M.E. (1982). Pseudodiagnosticity in an idealized medical problem-solving environment. *Journal of Medical Education*, *57*, 110–114.
Klayman, J. (1988). Cue discovery in probabilistic environments: uncertainty and experimentation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 317–330.
Klayman, J., & Ha, Y.-W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, *94*, 211–228.
Klayman, J., & Ha, Y.-W. (1989). Hypothesis testing in rule discovery: strategy, structure and content. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 596–604.
Komatsu, L.K. (1992). Recent views of conceptual structure. *Psychological Bulletin*, *112*, 500–526.
Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory*, *6*, 107–118.
Kriel, J., & A'Beckett-Hewson, M. (1986). Conceptual frameworks in preclinical and clinical textbooks. *Medical Education*, *20*, 94–101.
Kyburg, H.E., Jr (1983). Rational belief. *Behavioral and Brain Sciences*, *6*, 231–273.
Lee, A.Y., & Pennington, N. (1992, November). *Across-domain transfer from program debugging to*

electronic troubleshooting. Paper presented at the meeting of the Psychonomic Society, St. Louis.
Lemieux, M., & Bordage, G. (1992). Propositional versus structural semantic analyses of medical diagnostic thinking. *Cognitive Science*, *16*, 185–204.
McKenzie, C.R.M. (in press). The accuracy of intuitive judgment strategies: covariation assessment and Bayesian inference. *Cognitive Psychology*.
Medin, D.L., & Smith, E.E. (1984). Concepts and concept formation. *Annual Review of Psychology*, *35*, 113–138.
Nakao, M.A., & Axelrod, S. (1983). Numbers are better than words: verbal specifications of frequency have no place in medicine. *American Journal of Medicine*, *74*, 1061–1065.
Neufeld, V.R., Norman, G.R., Barrows, H.S., & Feightner, J.W. (1981). Clinical problem-solving of medical students: a longitudinal and cross-sectional analysis. *Medical Education*, *15*, 26–32.
Nisbett, R.E., Krantz, D.H., Jepson, C., & Kunda, Z. (1983). The use of statistics in everyday inductive reasoning. *Psychological Review*, *90*, 339–363.
Norman, G.R. (1988). Problem-solving skills, solving problems and problem-based learning. *Medical Education*, *22*, 279–286.
Norman, G.R., & Schmidt, H.G. (1992). The psychological basis of problem-based learning: a review of the evidence. *Academic Medicine*, *67*, 557–565.
Norman, G.R., Tugwell, P., Feightner, J.W., Muzzin, L., & Jacoby, L. (1985). Knowledge and clinical problem solving. *Medical Education*, *19*, 344–356.
Oden, G.C. (1987). Concept, knowledge, and thought. *Annual Review of Psychology*, *38*, 203–228.
Osherson, D.N., Smith, E.E., Wilkie, O., López, A., & Shafir, E. (1990). Category-based induction. *Psychological Review*, *97*, 185–200.
Patel, V.L., Evans, D.A., & Kaufman, D.R. (1989). A cognitive framework for doctor–patient interaction. In D.A. Evans & V.L. Patel (Eds.), *Cognitive science in medicine: biomedical modeling* (pp. 257–312). Cambridge, MA: MIT Press.
Patel, V.L., Groen, G.J., & Norman, G.R. (1991). Effects of conventional and problem-based medical curricula on problem solving. *Academic Medicine*, *66*, 380–389.
Payne, J.W., Bettman, J.R., & Johnson, E.J. (1990). The adaptive decision maker: effort and accuracy in choice. In R. Hogarth (Ed.), *Insights in decision making: a tribute to Hillel J. Einhorn* (pp. 129–153). Chicago: University of Chicago Press.
Politser, P.E. (1987). Medical education for a changing future: new concepts for revising texts. *Medical Education*, *21*, 320–333.
Rosch, E., & Mervis, C.B. (1975). Family resemblances: studies in the internal structure of categories. *Cognitive Science*, *7*, 573–605.
Russo, J.E. (1977). The value of unit price information. *Journal of Marketing Research*, *14*, 193–201.
Saks, M.J., & Kidd, R.F. (1980). Human information processing and adjudication: trial by heuristics. *Law and Society Review*, *15*, 123–160.
Schmidt, H.G., Dauphinee, W.D., & Patel, V.L. (1987). Comparing the effects of problem-based and conventional curricula in an international sample. *Journal of Medical Education*, *62*, 305–315.
Schoemaker, P.J.H., & Russo, J.E. (1992). *A pyramid of decision approaches* (working paper no. 246). Chicago: University of Chicago Graduate School of Business, Center for Decision Research.
Schwartz, S., & Griffin, T. (1986). *Medical thinking: the psychology of medical judgment and decision making*. New York: Springer.
Schwartz, W.B., Gorry, G.A., Kassirer, J.P., & Essig, A. (1973). Decision analysis and clinical judgment. *American Journal of Medicine*, *55*, 459–472.
Sherman, S.J., & Corty, E. (1984). Cognitive heuristics. In R.S. Wyer & T.K. Srull (Eds.), *Handbook of social cognition* (Vol. 1, pp. 189–286). Hillsdale, NJ: Erlbaum.
Slowiaczek, L.M., Klayman, J., Sherman, S.J., & Skov, R.B. (1992). Information selection and use in hypothesis testing: what is a good question, and what is a good answer? *Memory & Cognition*, *20*, 392–405.
Smith, J.F., & Kida, T. (1991). Heuristics and biases: expertise and task realism in auditing. *Psychological Bulletin*, *109*, 472–489.

Sox, H.C., Blatt, M.A., Higgins, M.C., & Marton, K.I. (1988). *Medical decision making*. Boston: Butterworths.

Tweney, R.D., Doherty, M.E., Worner, W.J., Pliske, D.B., Mynatt, C.R., Gross, K.A., & Arkkelin, D.L. (1980). Strategies of rule discovery in an inference task. *Quarterly Journal of Experimental Psychology*, *32*, 109–123.

von Winterfeldt, D., & Edwards, W. (1986). *Decision analysis and behavioral research*. Cambridge, UK: Cambridge University Press.

Wallsten, T.S., & Budescu, D.V. (1983). Encoding subjective probabilities: a psychological and psychometric review. *Management Science*, *29*, 151–173.

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

Weinstein, M.C., & Fineberg, H.V. (1980). *Clinical decision analysis*. Philadelphia: W.B. Saunders.

Wolf, F.M., Gruppen, L.D., & Billi, J.E. (1985). Differential diagnosis and the competing-hypothesis heuristic: a practical approach to judgment under uncertainty and Bayesian probability. *Journal of the American Medical Association*, *253*, 2858–2862.

Wortman, P.M. (1972). Medical diagnosis: an information processing approach. *Computational Biomedical Research*, *5*, 315–328.

Appendix A: summary chapter from the independent training condition (zymosis)

Identification: Zymosis is an acute disease of short duration and varying severity. Onset is sudden and characterized by fever, cough, prostration (physical weakness) and anorexia (loss of appetite). Anorexia is commonly associated with diarrhea, although the underlying cause is sometimes nausea. The degree of prostration is more likely to be mild than severe. Zymosis patients experience difficulty in walking; the cause is usually pain in the lower extremities, especially the feet. Sometimes, though, the patient's difficulty in walking is attributed to a "pins-and-needles" or tingling sensation in the legs and feet. Fever is constantly elevated for 5 to 7 days, and is very likely to be low-grade (i.e., temperatures in the range of 99–101°F). Patients normally develop a dry cough, although they occasionally experience a harsh, barking cough instead. More often than not, the tongue is cracked and fissured. As the disease progresses, a rash appears on the chest and abdomen and then spreads to the body generally. This rash commonly has a scaly consistency, although it sometimes can be moist and "weepy".

Occurrence: Worldwide and relatively frequent. More common in males than females, and somewhat more common in children than adults. There are pronounced racial differences, with White people being much more susceptible than Black or Brown people, for reasons that are not clear.

Mode of transmission: From person to person by direct contact. Highly contagious.

Incubation period: Highly variable and difficult to ascertain; usually 7 to 21 days.

Susceptibility and resistance: Susceptibility is universal among those not

previously infected. One infection confers long-term immunity. Immunity is sometimes conferred by subclinical infection (i.e., without the disease being obvious).

Appendix B: summary chapter from the contrastive training condition

Identification: Proxititis and zymosis are both acute diseases characterized by sudden onset of fever, cough, prostration (physical weakness) and anorexia (loss of appetite), and later development of a rash. Illness is of short duration, with constantly elevated temperature for about 5 to 7 days. The two diseases have similar symptomatology.

With zymosis, anorexia is commonly associated with diarrhea, although the underlying cause is sometimes nausea. The degree of prostration is more likely to be mild than severe. Patients experience difficulty in walking; the cause is usually pain in the lower extremities, especially the feet. Sometimes, though, the patient's difficulty in walking is attributed to a "pins-and-needles" or tingling sensation in the legs and feet. Fever is very likely to be low-grade (i.e., temperatures in the range of 99–101°F). Zymosis patients normally develop a dry cough, although they occasionally experience a harsh, barking cough instead. More often than not, the tongue is cracked and fissured. As the disease progresses, a rash appears on the chest and abdomen and then spreads to the body generally. This rash commonly has a scaly consistency, although it sometimes can be moist and "weepy".

With proxititis, fever is more often low-grade than high. Patients are somewhat more likely to have a dry cough than a harsh, barking one. The tongue is usually cracked and fissured but occasionally will have a smooth appearance. Anorexia is commonly due to nausea. Prostration is normally mild, although proxititis patients occasionally will experience more severe physical weakness. In addition, most patients experience difficulty in walking due to a tingling sensation in the lower extremities, particularly the feet. Later, a rash appears. The rash is usually moist and "weepy", although proxititis patients sometimes have a scaly rash.

Occurrence: Both diseases occur worldwide and are relatively frequent. Zymosis is more common in males than females, and somewhat more common in children than adults. There are pronounced racial differences, with White people being much more susceptible to zymosis than Black or Brown people, for reasons that are not clear. Proxititis is more common in females than in males. Racial differences are not pronounced, but Whites are somewhat more susceptible than Black and Brown people. Proxititis is normally a disease of children and adolescents, but is sometimes seen in adults.

Mode of transmission: From person to person by direct contact. Both diseases are highly contagious.

Incubation period: Highly variable and difficult to ascertain; usually 7 to 21 days for both diseases.

Susceptibility and resistance: Susceptibility is universal among those not previously infected. One infection confers long-term immunity. Immunity is sometimes conferred by subclinical infection (i.e., without the disease being obvious).