Risk Decision and Policy (2000), *vol.* 5, *pp.* 69–100. Printed in the United Kingdom

Confidence judgments as expressions of experienced decision conflict

ELKE U. WEBER*
Department of Psychology and Graduate School of Business, Columbia University, 406 Schermerhorn Hall, (MC 5501), New York NY 10027, USA
E-mail: euwz@columbia.edu

ULF BÖCKENHOLT
Department of Psychology, University of Illinois at Urbana-Champaign, 603 E. Daniel Street, Champaign, IL 61820, USA
E-mail: ubockenh@s.psych.uiuc.edu

DENIS J. HILTON
University of Toulouse

BRIAN WALLACE
University of Wales College of Medicine, 12 Holly Grove, Lisvane, Cardiff, Wales CF4 5UG

Abstract

This study tested between two interpretations of confidence in diagnostic hypotheses: expected probability of being correct and conflict experienced during the diagnostic process. Physicians generated hypotheses for case histories with two plausible diagnoses, one having a higher population base rate but less severe clinical consequences than the other. Case information indicative of the two diagnoses was varied. Generation proportions for the two diagnoses and confidence judgments both deviated from the predictions of a Bayesian belief model, but in different ways. Generation of a hypothesis increased with diagnosis-consistent information and diagnosis base rates, but was not reduced by diagnosis-inconsistent information. Confidence was sensitive to both consistent and inconsistent information, but was not very sensitive to diagnosis base rates. Physician characteristics also affected hypothesis generation and confidence differentially. Female doctors had lower confidence in their diagnoses than male doctors, yet there were no gender differences in hypothesis generation. Experience affected hypothesis generation monotonically via the increased availability of previously diagnosed cases, while confidence first increased and then decreased with doctors' experience. The results are consistent with an interpretation of confidence judgments as an expression of decision conflict rather than an indication of likely diagnosis accuracy.

1. Introduction

Confidence judgments have recently been the focus of renewed interest (e.g., Gigerenzer *et al.*, 1991; Griffin and Tversky, 1992; Soll, 1996), partly because of disagreement about their interpretation (Dawes and Mulford, 1996), and partly because of evidence suggesting that confidence helps to predict other observable behaviour (Heath and Tversky, 1991). Motivated by a scarcity of real-world studies

*Author to whom correspondence should be addressed.

in the judgment and decision literature (Robins and Craik, 1993), this study was designed to provide a better understanding of confidence in a real world setting involving diagnostic judgments made by individuals with experience in the content domain. In particular, we examined the determinants and consequences of the confidence that general practitioners hold in their medical diagnoses. In this context, we contrast two interpretations of confidence judgments. The first views confidence as an expression of subjective accuracy, i.e., the judge's subjective probability that his or her judgment is correct. The second views confidence as an expression of the conflict experienced during the process of generating the judgment.

2. Confidence

2.1. USES OF CONFIDENCE

In many real world decision situations, information about the objective accuracy of one's judgments is not readily available. Is a juror correct in having reached a 'guilty as charged' verdict? Is a physician correct in her working hypothesis about a patient? In such situations, the accuracy of the judgment is not only difficult to assess at the time of judgment – if it were, there would be no need for the expert judgment, but often it may not become known later on either, making it difficult to learn from experience. When no objective guides to accuracy are available, decision makers have to use some subjective impression of likely decision quality, usually their confidence in their judgment, as a guide to further action. Adams and Adams (1961) argued that a realistic level of confidence may in many situations be more important than the level of performance itself and defined 'realistic' in terms of long-run accuracy. Their paper started a long line of investigation on the calibration of confidence judgments that compares expressed confidence to the long-run probability of being correct (for a review see Yates, 1990, Chapter 4).

If people use their confidence in decisions or judgments as a guide to further action, it is important to understand the processes that give rise to the experienced level of confidence. Yet, most process-level investigations of confidence have focused on the determinants of *over*confidence (e.g., Oskamp, 1965; Hoch, 1985), rather than on the processes that generate confidence in the first place. For some tasks (e.g., weather forecasting), the appropriate absolute level of confidence is of great importance. For other purposes it may, however, be just as important to know what factors determine the 'relative' confidence somebody has in different courses of action than to know by how much these confidence judgments all exceed 'appropriate' accuracy levels, i.e., are poorly calibrated. Beach (1975) argued, for example, that in medical diagnosis the primary objective is to reach the correct diagnosis rather than to attach the right absolute magnitude of confidence to this diagnosis. If one is concerned only with the ordinal properties of confidence judgments, the whole issue of over- versus underconfidence, in fact, disappears. With the exception of Peterson and Pitz (1988), there is little work on confidence in judgments or decisions under uncertainty as a function of the type or amount of available information. Thus, without any intent to diminish the importance of research on calibration, this study deals with issues related to confidence that have so far received less attention.

2.2. CONFIDENCE AS ASSESSMENT OF ACCURACY

The currently popular interpretation of confidence judgments sees them as an expression of the likely accuracy of a judgment or decision. Confidence in a hypothesis, for example, is assumed to reflect one's assessment of the subjective probability that the hypothesis is correct, given the available data or evidence. Within this Bayesian framework, people may assess this subjective probability using different indicators, include their knowledge about the predictive validity of available cue information (Gigerenzer *et al.*, 1991), knowledge of the base rate with which an answer is correct (Bar-Hillel, 1980), the amount of evidence supporting an answer (Koriat *et al.*, 1980), or the perceived ease or speed with which a hypothesis or solution has come to mind (Schwarz *et al.*, 1991). Some of these indicators may be fallible and thus result in inaccurate estimates of likely accuracy (e.g., Oskamp, 1965). Nevertheless, all of these accounts assume that people intend to express the likely accuracy of their answer or estimate with their confidence judgment.

2.3. CONFIDENCE AS EXPRESSION OF DECISION CONFLICT

A more historical perspective shows that confidence judgments as a dependent variable have had a history in psychology that predates their interpretation in terms of calibration and accuracy. Looking for a graded measure of uncertainty in discrimination judgments, the early American psychophysicists Peirce and Jastrow (1884) discovered a simple functional relationship between average confidence judgments and the proportions of 'equal' as opposed to 'different' responses in psychophysical discrimination tasks under different conditions.[1] Response proportions (measured by the experimenter over trials) and confidence in the discriminations (expressed by the judge on each trial) were found to be equivalent and substitutable expressions of an individual's discrimination ability. Confidence judgments had the advantage of being easier to obtain than response proportions, providing reliable estimates of the relative difficulty of judgments after a small number of trials. Similar ease-of-acquisition arguments subsequently led to the use of confidence judgments in signal detection theory, where complete ROC curves are constructed on the basis of different degrees of confidence in a judgment (Clarke, 1960).

In these psychophysical tasks, confidence was seen as an expression of the subjective experience of the discrimination process, providing information about the difficulty experienced in arriving at the final decision. Instead of relating confidence to characteristics about the outcome of the decision (e.g., its accuracy), subsequent researchers related it to other process measures, such as response time. Henmon (1911), for example, showed that with discriminability held constant, judgments that had a slower response time were made with less confidence, suggesting that observers based their confidence judgments not on discriminability *per se*, but on process features such as the time taken to reach a decision. The interpretation of confidence as a reflection of characteristics of the decision 'process' in these early

[1] Confidence was proportionate to the log of the ratio of the proportion of 'equal' over the proportion of 'different' judgments in the comparison of simple unidimensional physical stimuli.

psychophysical experiments may hold some lessons for its interpretation in other decision situations.

There is some evidence to suggest that confidence judgments express the conflict experienced between one's belief in different answers or between one's preferences for different alternatives, in ways that go beyond the predictions of a Bayesian belief updating model (Janis and Mann, 1968; Peterson and Pitz, 1988; Paese and Sniezek, 1991). In situations where groups arrive at decisions by methods that either confront or do not confront conflict, induced by differences in opinion by group members, confidence in the group answer tends to be higher when confrontation of the conflict is avoided (Sniezek and Henry, 1989). Zakay (1985) found an even more direct relationship between confidence levels and decision conflict. In decisions made by nurses, ratings of their post-decision confidence were significantly higher for those choices made using noncompensatory processing entailing less decision conflict, than for those choices made using a compensatory strategy that necessitated conflictual tradeoffs. These accounts see confidence as reflecting characteristics of the *process* that gives rise to the judgment or decision. The nature of the available information as well as context and task features all affect that process, facilitating or complicating it in a number of ways. In this interpretation, the experienced level of confidence in the final answer serves as 'memory' of the nature of the judgment or decision process that gave rise to it.

2.4. PRESENT STUDY

Medical diagnosis is an important domain in which to study the origins and consequences of judgments of confidence. Clinical problems are often without clearly correct solutions, necessitating the use of subjective confidence as a proxy for likely accuracy. A substantial body of evidence demonstrates that physicians generate a number of diagnostic hypotheses, i.e., working interpretations of their patient's presenting complaints, very early in the consultation (e.g., Elstein *et al.*, 1978; Barrows *et al.*, 1982). The composition of this set of hypotheses can vary considerably. In a family practice setting, the great majority of cases seen by a physician are either common and benign routine diagnoses or involve the management of chronic (known) diseases. However, a small fraction of cases are medical problems with consequences that are serious (e.g., life-threatening) if they go undetected. In analogy to a signal detection problem, the challenge for the family practitioner is to detect those cases (the signals) among the noise of routine cases.

To examine the susceptibility of confidence judgments to conflict between competing response alternatives, we designed clinical scenarios that contained some interpretative ambiguity between common-and-benign versus rare-but-serious diagnoses and varied the symptom information indicative of one or the other or both diagnoses. Practicing physicians generated diagnostic hypothesis for each case. They also rated their confidence in the complete set of generated hypotheses as well as their confidence in the two hypotheses they judged first- and second-most likely. The experimental design allowed us to look at the effects of three determinants of likely accuracy of a particular diagnosis on hypothesis generation and confidence judgments: (a) the population base rate of the diagnosis, (b) the amount of symptom information consistent with the diagnosis, and (c) the amount of symptom information inconsistent with the diagnosis. Previous work on the relationship between amount of information and confidence (Oskamp, 1965; Peterson and Pitz, 1988) did not distinguish between information that was consistent as opposed to inconsistent with the provided judgment and thus had no opportunity to explore conflict interpretations of confidence. These and other studies (Lichtenstein and Fischhoff, 1977; Griffin and Tversky, 1992) looked at the effect of some information variables on confidence, but did not contrast these effects with the effect of such variables on hypothesis generation itself. Other determinants of confidence – e.g., the credibility of cue information (Griffin and Tversky, 1992) or the validity of diagnostic cues (Gigerenzer *et al.*, 1991) – were held constant in our study in order to observe more clearly any effects of the three variables listed above.

Characteristics of the diagnostician may also affect confidence, but have been largely ignored in past investigations. Few empirical studies of confidence have systematically investigated the effects of expertise or experience on confidence. Those that did (Goldberg, 1959; Oskamp, 1962, 1965) used a fairly narrow range of experience and focused primarily on changes in *over*confidence, which was found to decrease with experience. The present study was designed to determine how individual differences between diagnosticians affect the processes involved in hypothesis generation and confidence assessment.

2.5. JUDGMENTS OF CONFIDENCE AND THEIR CONSEQUENCES

We were interested to identify what exactly physicians were confident about, and how their confidence related to hypothesis generation and subsequent diagnostic or treatment actions. Elstein *et al.* (1978) showed that physicians frequently consider several diagnoses simultaneously when making diagnoses. Our own pilot observations indicated that family practitioners often felt confident that one of the diagnoses they generated would be correct, without having enough information initially to decide which one it would be. Protracted ambiguity of this sort is relatively prevalent in a domain such as family medicine, where disease management decisions often have to be made at times when a full range of symptoms is not yet available, due to disease processes having yet to run their course or due to time lags in obtaining test or treatment results.

Thus, we were interested in how physicians' confidence in their overall set of diagnoses related to their confidence in individual diagnoses. Were physicians more confident that their overall set of diagnoses would be correct when it had one clear 'winner?' Or would they be more confident that the overall set was correct when it contained two or more likely 'winners?' Note that from a Bayesian point of view, physicians should be more confident in their overall set in the second case. However, a conflict model of confidence might predict less confidence in the second case.

In addition, we wanted to know which measure of confidence had the most validity in predicting subsequent aspects of a dynamic decision-making process. For example, does confidence in the set of diagnoses or in one of the individual diagnoses best predict physicians' decisions to collect further information? Böckenholt *et al.* (1991) found that differences in the relative attractiveness of alternatives determined the amount of information that people acquired about the alternatives prior to choice, possibly mediated by differences in confidence in their choice. To test this hypothesis, we asked physicians questions about the follow-up information and test results that they would likely request before deciding on their final diagnosis, in order to address the question of what subsequent behaviours were affected by the different confidence judgments.

3. Method

3.1. SUBJECTS

Participants were family practitioners in South Glamorgan, Wales, UK, working in group practices which varied in size from 2 to 8, with a mean doctor/patient ratio of 1:2000. Approximately 30% of practices are recognized as training practices for the postgraduate education of trainee general practitioners (GPs). All doctors working in training practices in this district were asked to participate.

Contacted physicians were either trainers (fully registered GPs recognized as teachers of trainee GPs; $N = 53$), nontrainers (fully registered GPs who were partners of trainers in training practices; $N = 41$), or trainees (recently registered medical practitioners undergoing a 12 months postgraduate training programme to become GPs; $N = 37$). The response rate was 64.1% (similar for all three categories of respondents), for a total of 84 respondents (58 male and 26 female). Experience varied from one month past medical school to 38 years of clinical practice.

3.2. CASE HISTORIES

The three case histories used in the study were based on real patients whose names and personal details were changed to avoid identification. Two pilot studies established that physicians would accept the case histories as realistic and determined plausible diagnostic hypotheses. Based on the results of the pilot studies, the physician on our research team in consultation with another experienced physician identified two plausible diagnostic hypotheses (A and B) for each case history. These diagnoses were, for all practical purposes, mutually exclusive (i.e., it was highly unlikely that a patient would suffer from both) and asked for different treatments. In all three cases, medical reference books showed the A-diagnosis as having a higher population base rate but less severe clinical consequences than the B-diagnosis. Weber *et al.* (1993) report several manipulation checks, validating that doctors perceive these differences between the two types of diagnoses correctly. Other diagnoses (O) were also conceivable, but not very plausible given the symptomatic and background information. For Case I, for example, the A-diagnosis was 'upper gastrointestinal disease', the B-diagnosis 'ischaemic heart disease', and O-diagnosis included 'lung condition', 'anxiety', or 'gall bladder disease'.[2]

Using the results of the pilot studies, 16 versions of each case history were created, where each version had a different combination of clinical symptoms and background information as described below. The most complete version for Case I read as follows:

[2] The A, B, and O hypotheses for all three casees, together with a listing of the specific diagnostic labels generated by our respondents are available upon request.

> Mr. Brooks is 45 years old, married with two daughters, and works as a long distance lorry driver. He has had a burning lower retrosternal discomfort while driving for three months. In the past he has been troubled by recurrent dyspepsia and frequent chest infections. He smokes 40 cigarettes a day. His father died, age 52, following a myocardiac infarction.

In contrast, the most reduced version read as follows:

> Mr. Brooks is 25 years old, married with two daughters, and works as a postman. He has had a lower retrosternal discomfort for three months. In the past he has been troubled by frequent chest infections. He smokes 40 cigarettes a day.

3.3. DESIGN AND PROCEDURE

The experimental manipulation of information followed a 2 × 2 × 2 × 2 factorial design that varied the amount of clinical and background information indicative of diagnosis A and the amount of clinical and background information indicative of diagnosis B that doctors received, with two levels (full versus reduced) for each of the four information factors. Table 1 illustrates the design by listing the clinical and background information indicative of diagnosis A and B, respectively, that doctors received under the full and reduced conditions for Case I.[3] By crossing the four clinical conditions shown there (full versus reduced clinical information indicative of A and/or B) with the four background information conditions (full versus reduced background information indicative of A and/or B), 16 different versions of each case history were generated that were presented to physicians in a between-subject design.

Doctors answered a nine-page paper-and-pencil questionnaire that contained one randomly selected version of each of the three cases. For each case history, they answered the following set of questions: (1) What could be wrong with this patient? Please list as many possibilities as you would consider in real life. If there is more than one, please list them in order of likelihood. (2) How confident are you at this stage that at least one of your ideas will be correct? (3) Which of your ideas would you explore first? Why have you chosen this one? (4) Have you encountered a similar problem before in practice? If so, what was the diagnosis on that occasion? Physicians were then asked to consider the diagnosis that they had listed first as most likely and to answer the following, additional, questions about it: (5) Please identify which items of the information provided first made you think of this idea? (6) How confident are you in making this diagnosis? What factors have led you to make this confidence rating? (7) What additional specific items of information would you require before you would decide upon this as a working diagnosis? They were then prompted to consider the diagnosis that they had listed in second place and to answer questions (5) through (7) again, now in reference to their second hypothesis. Confidence ratings were made by placing a mark on a 4-inch long graphic rating scale that ranged from 'Not at all Confident' to 'Extremely Confident'. These graphic line markings were subsequently recoded to a numerical rating ranging from 0 to 10.

The questionnaire was mailed to the GPs in our sample together with a cover letter

[3] A listing of the equivalent conditions for the other two cases can be found in Table 2 of Weber *et al.* (1993).

Table 1. Clinical and background information indicative of diagnoses A and B provided in the different information conditions for case I. Diagnosis A: upper gastro-intestinal disease and Diagnosis B: ischaemic heart disease.

	Information condition			
	reduced for A, *reduced for B*	*full for A,* *reduced for B*	*reduced for A,* *full for B*	*full for A,* *full for B*
Clinical information				

Common	Lower retrosternal discomfort;
	For 3 months;
	Past chest infections;
	Smokes 40 cigarettes/day
		Lower retrosternal discomfort;
		For 3 months;
		Past chest infections;
		Smokes 40 cigarettes/day
			Lower retrosternal discomfort;
			For 3 months;
			Past chest infections;
			Smokes 40 cigarettes/day
				Lower retrosternal discomfort;
				For 3 months;
				Past chest infections;
				Smokes 40 cigarettes/day

Indicative of A		Burning; Past recurrent dyspepsia		Burning; Past recurrent dyspepsia
Indicative of B			While driving; Father died age 52 of myocardial infarction	While driving; Father died age 52 of myocardial infarction
Background information				
Common	Gender: Male; Married, two daughters	Gender: Male; Married, two daughters	Gender: Male; Married, two daughters	Gender: Male; Married, two daughters
Indicative of A	Occupation: postman	Occupation: long distance truck driver	Occupation: postman	Occupation: long distance truck driver
Indicative of B	Age: 25	Age: 25	Age: 45	Age: 45

from the fourth author, a senior lecturer in the Department of General Practice at the University of Wales College of Medicine, asking them for their voluntary participation. Physicians were instructed to answer the questions for each case as if the patient described in the case history had just walked in for a consultation. Participants returned the completed questionnaire by mail. While responses were anonymous, the coding of provided return envelopes allowed for the identification of participants who had failed to return their questionnaire after some period of time. These physicians received a follow-up phone call asking them for their participation a second time. Doctors were assured that their responses would not be used to evaluate them in any way and were promised a summary of the results of the study.

4. Results

Doctors generated a list of hypotheses, rank-ordered it perceived likelihood, and provided three ratings of confidence: overall confidence that their generated set of hypotheses contained the correct diagnosis, confidence in the hypothesis they considered most likely, and confidence in the hypothesis they considered second-most likely. Analysis of the rank order of these three ratings showed that, in virtually every case and for every physician, the rating of confidence in the set was larger than confidence in the first-ranked hypothesis, which in turn was larger than confidence in the second-ranked hypothesis, suggesting that doctors used the rating scales sensibly and consistently.

Figure 1 shows the frequency of A-, B-, and O-diagnoses that physicians generated with different degrees of confidence as the first- and secondmost-likely hypothesis, respectively.[4] It is easy to see that, summed across the 16 information conditions, physicians generated more A-diagnoses than B-diagnoses, especially as their most-likely hypothesis. These joint distributions of the particular diagnoses doctors generated and their associated confidence ratings (f(diagnosis, confidence)) can be decomposed into the distribution of generated diagnoses (g(diagnosis)) and the distribution of confidence judgments given each diagnosis (h(confidence|diagnosis)), i.e.,

$$f(\text{diagnosis, confidence}) = g(\text{diagnosis}) \times h(\text{confidence}|\text{diagnosis}) \tag{1}$$

The generation proportions of different diagnoses and the degree of confidence given each diagnosis were both analysed as a function of the experimental information manipulations. For simplicity of exposition and to obtain reliable point estimates, we analysed the effect of variation in clinical and background information separately, i.e., we averaged the generation proportions and confidence judgments across the four different background information conditions for each of the four clinical information conditions and vice versa. The resulting data points are shown in Table 2. Both dependent measures were also compared to the predictions of a Bayesian belief model.

[4] To simplify exposition, O-diagnoses will not be mentioned explicitly in subsequent analyses. Their generation proportions and confidence judgments were similar to those of B-diagnoses and are available on request.

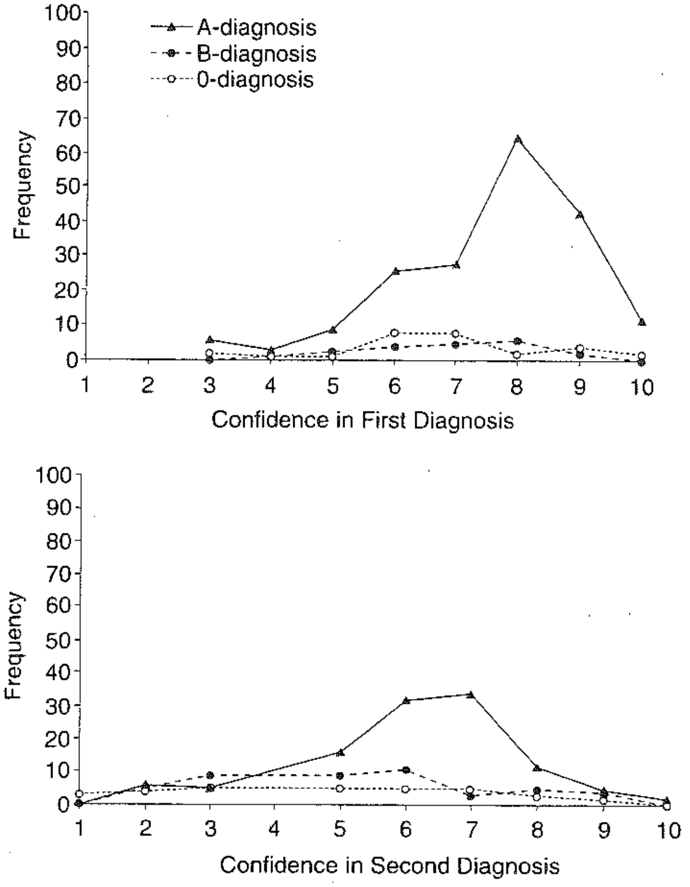

Figure 1 Frequency of A, B, and O diagnoses generated as the first- and second most likely diagnostic hypothesis, respectively, at each level of confidence.

4.1. QUALITATIVE BAYESIAN BELIEF PREDICTIONS

A Bayesian belief model that incorporates diagnosis base rates and the diagnosticity of the symptoms presented to physicians in different cells of the design allows computation of a normative estimate of the probability that a particular diagnosis (A or B) is correct, given the evidence. These normative estimates can then be compared against two dependent measures. The first measure is the proportion of physicians who judged the diagnosis to be the most (or second-most) likely diagnosis. Since the physicians in our sample had expertise in their diagnostic judgments and since the clinical cases were representative of the problems commonly solved by them, one might expect that the frequency with which they listed A- or B-type diagnoses as most (or second-most) likely should be related to the Bayesian probability that these diagnoses are correct given the case information and the diagnosis base rates. The second measure is the confidence that physicians expressed in the diagnosis that they generated as most (or second-most) likely. If confidence is an expression of doctors' subjective feeling of the likely accuracy of their diagnosis, then confidence judgments should also be related to the Bayesian probability of the diagnosis being correct. If both hypothesis generation and

Table 2. Proportions of generated diagnoses that were of an A- or B-type and conditional mean confidence judgments for A- and B-diagnoses as a function of amount (f = full; r = reduced) of clinical (Cl) and background (B) information indicative of A and B. Part (a) shows the values for the diagnoses doctors generated first. Part (b) shows the corresponding values for the diagnoses doctors generated second.

			Generation proportion		*Confidence given diagnosis*	
(a)						
Cl_A	Cl_B	*N*	A	B	A	B
f	f	75	0.85	0.11	6.9	6.6
f	r	66	0.96	0.02	7.7	5.1
r	f	63	0.62	0.14	5.9	6.3
r	r	47	0.68	0.06	6.7	7.9
B_A	B_B	*N*	A	B	A	B
f	f	60	0.83	0.10	6.9	5.8
f	r	72	0.86	0.08	6.9	6.1
r	f	60	0.76	0.10	6.7	7.1
r	r	59	0.69	0.05	7.4	6.4
(b)						
Cl_A	Cl_B	*N*	A	B	A	B
f	f	71	0.49	0.34	5.2	4.4
f	r	62	0.58	0.27	5.9	3.8
r	f	62	0.51	0.29	4.8	5.1
r	r	43	0.65	0.11	5.7	5.2
B_A	B_B	*N*	A	B	A	B
f	f	59	0.54	0.36	5.3	4.3
f	r	69	0.54	0.22	5.6	4.4
r	f	58	0.53	0.33	5.1	4.3
r	r	52	0.59	0.17	5.6	5.0

confidence judgments agree with the Bayesian predictions, then both measures are substitutable expressions of the likely accuracy of the diagnosis. If, on the other hand, confidence judgments deviate from the Bayesian estimates or if generation proportions and confidence judgments deviate from the Bayesian estimates in different ways, confidence becomes a measure that is useful in its own right, by expressing information different from likely accuracy that is not being expressed by the generation proportions.

What predictions does a Bayesian updating model make for the information conditions encountered by diagnosticians in our study? Figure 2 builds up the simplest set of qualitative predictions for the information manipulations of our study. The predictions are for the probability that an A-diagnosis (left panels) or a B-diagnosis (right panels) is correct, given the base rate of A- and B-diagnoses and the amount of A- and B-information provided. If generation proportions of diagnoses and/or confidence in diagnoses are related to the probability of being correct, the

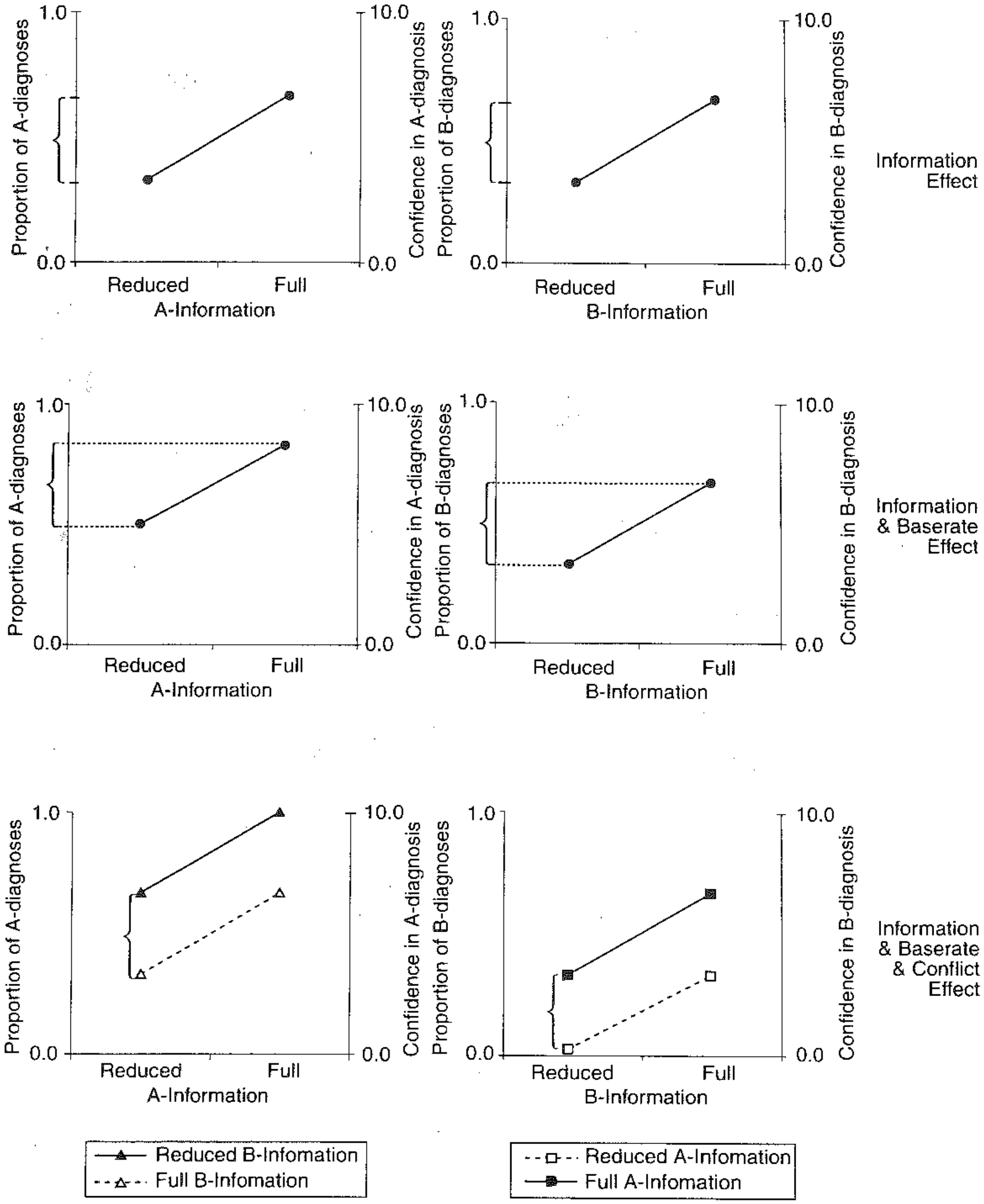

Figure 2 Predictions of a Bayesian belief updating model for the probability that an A- or B-diagnosis, respectively, is correct, given different amounts of clinical and background information indicative of an A or B diagnosis.

Bayesian probabilities will predict at least the rank order across information conditions of these two dependent measures.

As shown in the top panels, the probability of an A-diagnosis being accurate increases when, *ceteris paribus*, more clinical information indicative of an A-diagnosis ('full' rather than 'reduced' levels) is provided. A corresponding information effect is predicted to occur for the probability of a B-diagnosis being accurate, i.e., an increase when 'full' rather than 'reduced' levels of clinical information indicative of B are provided. The base rate effects predicted by the Bayesian model are shown in the middle set of panels: Given that A-diagnoses have a greater base rate of occurrence than B-diagnoses, the probability that an A-diagnosis is correct will be greater than the probability of a B-diagnosis, all other things being equal. Figure 2 shows the prediction for such a base rate effect as an additional main effect (i.e., as a shift in intercept), superimposed on the information effect. However, more complex patterns due to interactions between the two effects are also possible. Finally the bottom row of panels shows the Bayesian predictions for the effect of conflicting information: The probability that an A-diagnosis is accurate is lower when there is more evidence indicative of a competing diagnosis, i.e., when 'full' rather than 'reduced' levels of clinical information indicative of a B-diagnosis are present (bottom left panel). The corresponding prediction holds for the probability that a B-diagnosis is correct, i.e., a reduction when 'full' rather than 'reduced' levels of clinical information indicative of an A-diagnosis are present (bottom right panel). Again, consequences of a conflict effect are shown as an additional main effect, but two- or three-way interactions with the other two effects are also possible, as illustrated below.

4.2. EFFECTS OF INFORMATION ON GENERATION PROPORTIONS

Figure 3 shows the proportions with which physicians generated A-diagnoses (left panels) and B-diagnoses (right panels) as most-likely.[5] Testing for an effect of consistent information on the generation of an A-diagnosis involves the comparison of generation proportions under the two full versus the two reduced level conditions for information indicative of A, i.e., testing for a positive slope of the lines connecting the reduced and the full A-information conditions in the two left panels of Figure 3. An information effect in the generation of a B-diagnosis correspondingly involves the comparison of generation proportions under the two full versus the two reduced level conditions for information indicative of B, i.e., tesing for a positive slope of the lines connecting the reduced and the full B-information conditions in the two right panels. Generation frequencies were compared by chi-square tests under the null-hypothesis of equality-of-proportions. Five out of the eight possible tests showed a statistically significant information effect.[6] Generation proportions for a diagnosis increased as more information consistent with this diagnosis was provided for clinical and background A-information and clinical B-information for the diagnosis generated as most likely,

[5] Generation proportions for the second-most likely diagnosis looked similar (see Table 2). The precise numerical values of the generation proportions and of the confidence judgments, together with the sample size on which each data point is based, are shown in Table 2.

[6] Statistical significance for all reported results is at least at the 0.05 level.

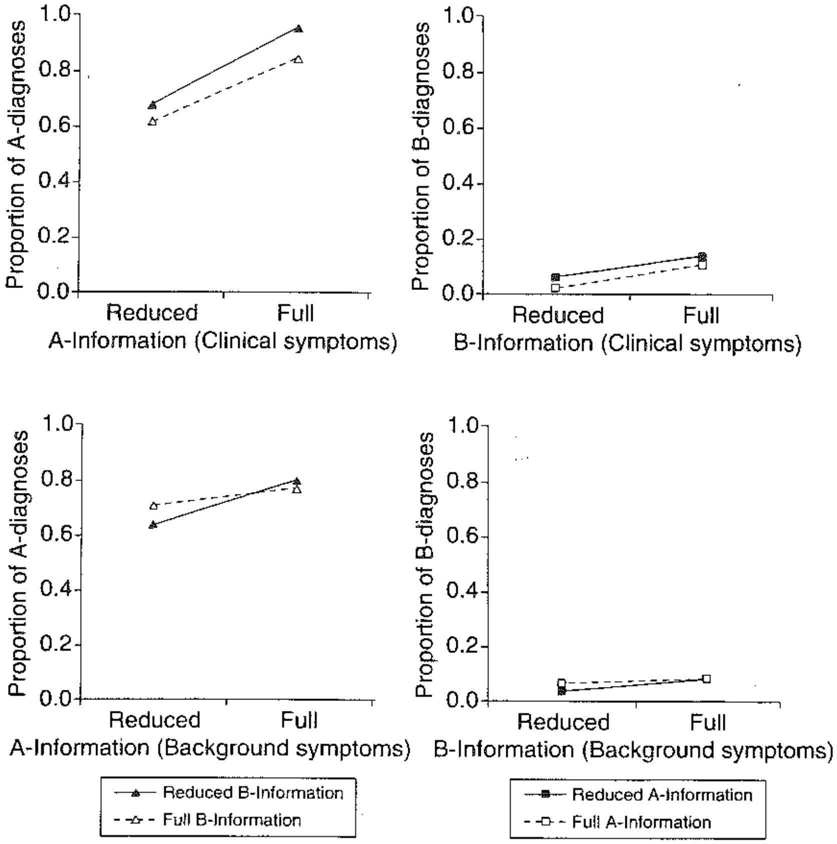

Figure 3 Observed proportions of A- and B-diagnosis, respectively, generated by physicians as their most-likely hypothesis, as a function of amount of clinical and background information.

and for clinical and background B-information for the diagnosis generated as second-most likely.[7]

Testing for a conflict effect of inconsistent information on the generation of an A-diagnosis involves the comparison of generation frequencies under the full versus the reduced conditions for diagnosis-inconsistent B-information, with the hypothesis that more inconsistent information will reduce the generation frequency of A-diagnoses. In Figure 3 a conflict effect exists when the solid line (for reduced inconsistent information) has a higher intercept than the dotted line (for full inconsistent information), i.e., when there is a gap between those two lines. Only one out of the eight possible tests showed a statistically significant conflict effect. In particular, more inconsistent clinical A-information reduced the generation proportions of B-diagnoses provided by physicians as their second-most likely hypothesis.

Testing for an effect of the differential base rates of A- versus B-diagnoses involves a comparison between corresponding information conditions in the right and left

[7] For simplicity and clarity of presentation, in this and subsequent analyses, results are reported across cases unless there were qualitative differences between cases, in which case the results of the by-case analyses are reported.

panels of Figure 3. In other words, a base rate effect occurs when the greater-base rate A-diagnosis is generated more frequently than the lower-base rate B-diagnosis. Generation frequencies for A- versus B-diagnoses showed a significant base rate effect for both physicians' first and second diagnosis. Across all conditions, A-diagnoses were generated significantly more frequently than B-diagnoses.

4.3. EFFECTS OF INFORMATION ON CONFIDENCE

The mean judgments of confidence in A- and B-diagnoses for the first diagnosis shown in Figure 4 show a very different pattern than the generation proportions in Figure 3. Confidence judgments were analysed for effects of the information conditions using ordinal regression[8] under the (conservative) assumption that confidence judgments may communicate only rank-order information (e.g., doctors may not have used the rating scale in a linear fashion).

The effect of consistent information on confidence in an A-diagnosis (B-diagnosis) predicted by Bayesian updating would result in differences in confidence between

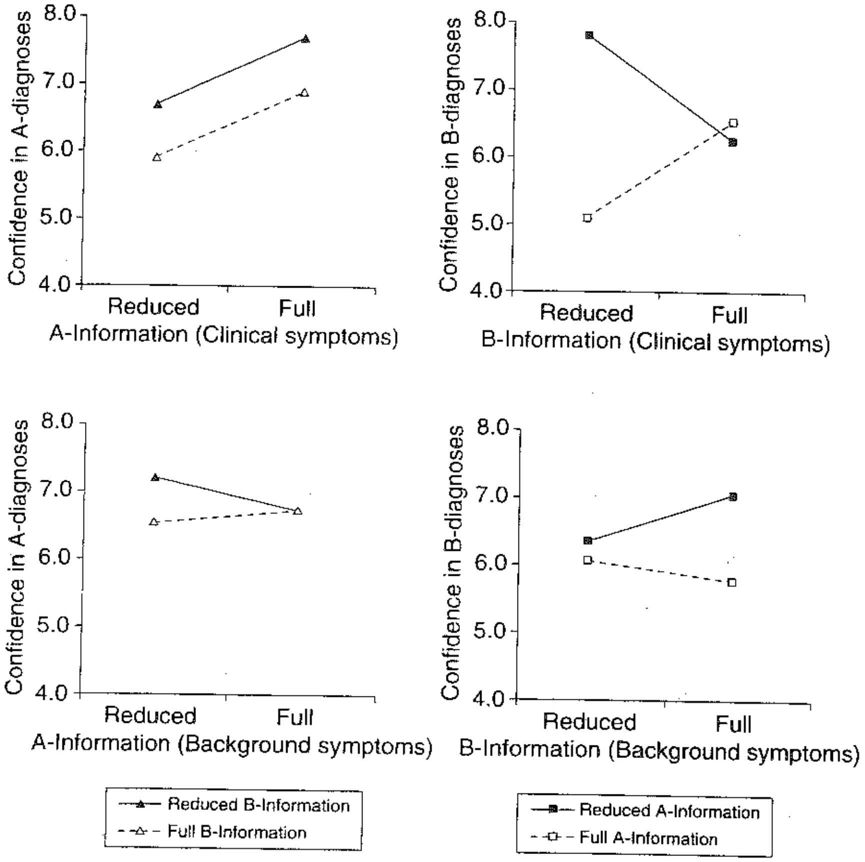

Figure 4 Observed mean confidence in A- and B-diagnosis, respectively, generated by physicians as their most-likely hypothesis, as a function of amount of clinical and background information.

[8] Full versus reduced levels of the four information factors were dummy coded with '1' denoting full information and '0' reduced information.

the two 'full' versus the two 'reduced' level conditions for information indicative of A (B), i.e., in a positive slope of the lines connecting the reduced to full A-information conditions in the two left panels (right panels) of Figure 4. Only one out of the eight possible tests showed an information effect. Except for confidence in A-diagnoses provided as the most-likely diagnosis, which increased as more clinical information consistent with A was provided (top left panel of Figure 4), confidence in either A- or B-diagnoses did not change as a function of amount of consistent information.

'Conflict' effects of inconsistent information on the other hand (that were largely absent in the generation proportions) were abundant for judgments of confidence. As before, a conflict effect results in a gap between the solid line (for 'reduced' inconsistent information) and the dotted line (for 'full' inconsistent information). Seven out of the eight possible tests showed conflict effects.

There was no significant effect of diagnosis base rates on doctors' confidence in either their first or second diagnosis, i.e., no difference between corresponding information conditions in the right and left panels of Figure 4 or the corresponding cells for the second-most-likely diagnosis.

4.4. RELATIONSHIP BETWEEN GENERATION PROPORTIONS AND CONFIDENCE JUDGMENTS

In summary, generation proportions were affected primarily by consistent information and diagnoses base rates, whereas confidence judgments showed evidence primarily of conflict effects brought about by inconsistent information. Confidence judgments and generation proportions differed to the point of often very different rank orders across the different information conditions.[9]

4.5. QUANTITATIVE BAYESIAN BELIEF PREDICTIONS

To see the extent and manner in which generation proportions and confidence judgments differed from Bayesian estimates, we computed Bayesian predictions for the effects of consistent and inconsistent clinical information and base rates in the following way. The conditional probability that an A-diagnosis or a B-diagnosis (respectively) is correct, given the information provided about symptoms indicative of A or B (I_A, I_B), is given by:

$$p(\mathrm{A}|I_\mathrm{A}, I_\mathrm{B}) = [p(I_\mathrm{A}, I_\mathrm{B}|\mathrm{A})\,p(\mathrm{A})]/[p(I_\mathrm{A}, I_\mathrm{B}|\mathrm{A})\,p(\mathrm{A}) + p(I_\mathrm{A}, I_\mathrm{B}|\mathrm{B})\,p(\mathrm{B}) + p(I_\mathrm{A}, I_\mathrm{B}|\mathrm{O})\,p(\mathrm{O})] \tag{2}$$

$$p(\mathrm{B}|I_\mathrm{A}, I_\mathrm{B}) = [p(I_\mathrm{A}, I_\mathrm{B}|\mathrm{B})\,p(\mathrm{B})]/[p(I_\mathrm{A}, I_\mathrm{B}|\mathrm{A})\,p(\mathrm{A}) + p(I_\mathrm{A}, I_\mathrm{B}|\mathrm{B})\,p(\mathrm{B}) + p(I_\mathrm{A}, I_\mathrm{B}|\mathrm{O})\,p(\mathrm{O})] \tag{3}$$

where $p(I_\mathrm{A}, I_\mathrm{B}|\mathrm{A})$ refers to the conditional probability of observing the particular symptom configuration (I_A, I_B) given that the A-disease is present. $p(\mathrm{A})$ is the population base rate of the A-disease. The other terms refer to corresponding

[9] The mean rank order correlation between the two dependent measures across the information conditions listed in Table 2 is 0.66 for A-diagnoses and −0.25 for B-diagnoses.

probabilities for B- and O-diseases. Two independent expert physicians provided us with approximate estimates of the base rates of A-, B-, and O-diagnoses for our three cases in the population with which the physicians in our sample were familiar. Estimates were quite similar for all three cases, and there was close agreement between the two experts: $p(\text{A}) = 0.7$, $p(\text{B}) = 0.1$, and $p(\text{O}) = 0.2$ (where $p(\text{O})$ is the sum of the base rates of a large number of 'other' diagnoses). Our experts also provided estimates of the conditional probabilities of particular sets of clinical symptoms given an A- or B-diagnosis. 'Reduced' symptom information (r) refers to the list of 'common' clinical symptoms shown in Table 1. 'Full' symptom information (f) included the additional symptoms listed there as indicative of either diagnosis A or B. Estimated conditional probabilities were again similar across cases: $p(r_{\text{A}}|\text{A}) = p(r_{\text{A}}|\text{B}) = p(r_{\text{A}}|\text{O}) = p(r_{\text{B}}|\text{A}) = p(r_{\text{B}}|\text{B}) = p(r_{\text{B}}|\text{O}) = 0.3$; $p(f_{\text{A}}|\text{A}) = 0.6$, $p(f_{\text{A}}|\text{B}) = 0.15$, $p(f_{\text{A}}|\text{O}) = 0.2$; $p(f_{\text{B}}|\text{A}) = 0.15$, $p(f_{\text{B}}|\text{B}) = 0.8$, $p(f_{\text{B}}|\text{O}) = 0.5$. Our experts approved the simplifying assumption that the conditional probability of information indicative of A and B given that disease A is true (i.e., $p(I_{\text{A}}, I_{\text{B}}|\text{A})$) is the product of the conditional probabilities of the two information components: $p(I_{\text{A}}, I_{\text{B}}|\text{A}) = [p(I_{\text{A}}|\text{A})][p(I_{\text{B}}|\text{A})]$.

The Bayesian predictions for the conditional probabilities of an A-diagnosis or a B-diagnosis given the four information conditions for clinical symptoms indicative of A or B ($(f_{\text{A}}, f_{\text{B}})$, $(f_{\text{A}}, r_{\text{B}})$, $(r_{\text{A}}, f_{\text{B}})$, $(r_{\text{A}}, r_{\text{B}})$), computed by using the estimates provided in the last paragraph, are shown in the top panels of Figure 5. These specific quantitative predictions confirm the qualitative predictions of Figure 2. Bayesian updating of base rates to incorporate available clinical information predicts an effect of consistent information (positive slope of lines), a conflict effect of inconsistent information (lower probabilities for the 'full' inconsistent than 'reduced' inconsistent information conditions), and a base rate effect (lower probabilities for the low base rate B-diagnoses (right panels) than high base rate A-diagnoses (left panel)).

Having specific predictions for the Bayesian probability correct of A- and B-diagnoses under the different information conditions allows us to compare these predictions to the observed response proportions and confidence judgments. The middle panels of Figure 5 show the comparison between observed generation proportions of A- and B-diagnoses (dotted lines) and the Bayesian predictions (solid lines). The bottom panels show the comparison between observed confidence in A- and B-diagnoses (dotted lines) and the Bayesian predictions (solid lines).[10] Figure 5 visually confirms the results of the statistical analyses in the previous section that showed that both dependent measures, generation proportions and confidence judgments, deviate from Bayesian estimates of the probability correct, but in different ways. Generation proportions reflect the effect of inconsistent information and diagnosis base rates appropriately, but fail to show the negative effect of inconsistent information predicted by the Bayesian model. Confidence judgments, on the other hand, coincide with the Bayesian predictions reasonably well for confidence in the high base rate A-diagnoses, but deviate drastically from the Bayesian predictions for the low base rate B-diagnoses (lower right panel of Figure 5). The effect of consistent information on confidence in a B-diagnosis seems to depend on whether or not

[10] Since confidence judgments and probabilities are on different scales, only the pattern of rank orders ought to be interpreted.

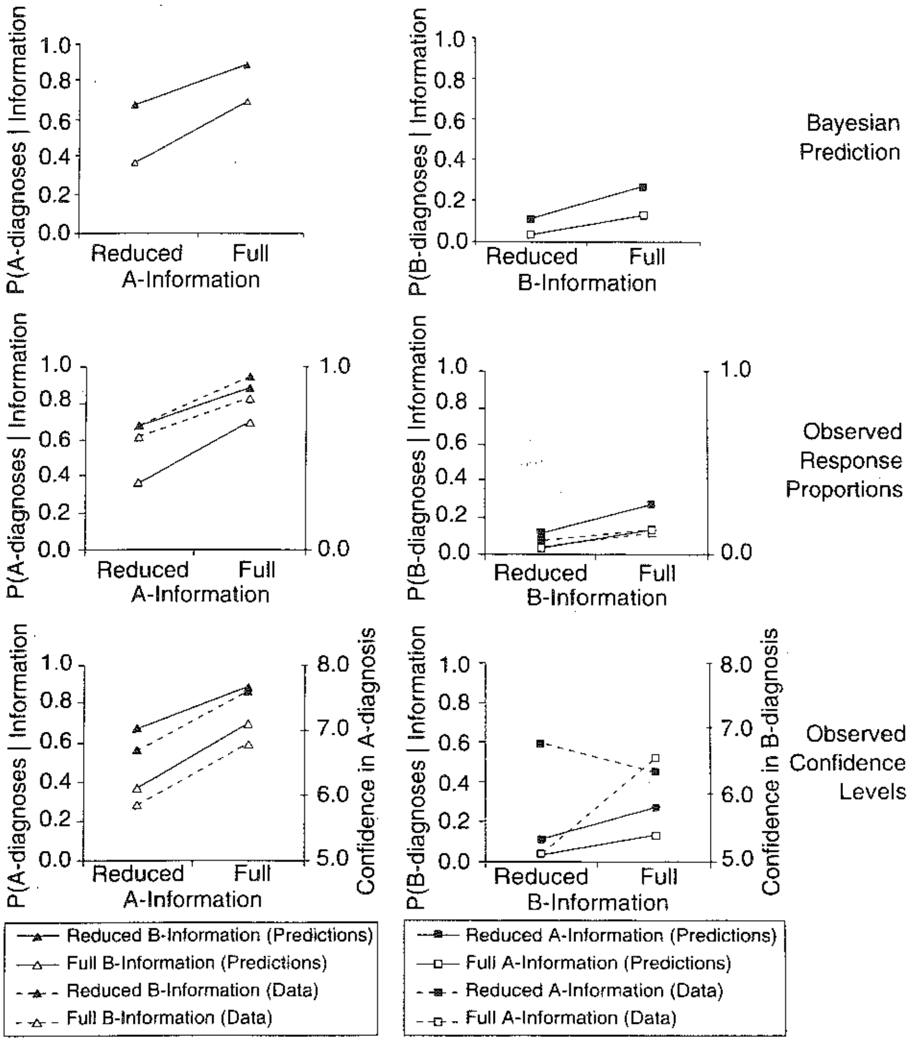

Figure 5 Bayesian predictions of the probability correct of an A-diagnosis (left panels) or B-diagnosis (right panels) for different clinical information conditions, and comparison between Bayesian predictions and observed generation proportions (middle panels) and observed confidence levels (bottom panels).

inconsistent information indicative of A is present. When inconsistent information is present (white squares), then the effect of consistent information ('full' versus 'reduced' levels of B-information) is much larger than predicted by the Bayesian model. When no inconsistent information is present (black squares), additional B-information actually has a small (but nonsignificant) negative impact on confidence in the B-diagnosis. Apparently the conflict induced by the presence of inconsistent information gets physicians to put greater weight on the amount of consistent information available when they make their confidence judgments, to the point of resulting in greater confidence in a B-diagnosis for the (full-A, full-B) than for the (reduced-A, full-B) information condition, in a between-subject design. It is perhaps not surprising that the conflict induced by inconsistent information is much stronger when inconsistent information indicative of a high base rate A-diagnosis is present in the face of tending towards a low base rate B-diagnosis, than vice versa. The point most different from the Bayesian probability correct prediction is the confidence in a B-diagnosis under reduced information conditions for both A and B. Confidence in the low base rate B-diagnosis is highest for this condition which provides the least amount of information that could either confirm or disconfirm the diagnosis, suggesting that, at least for B-diagnoses, conflict effects outweigh information effects for judgments of confidence.

4.6. CONFIDENCE IN SET OF DIAGNOSES

The previous section discussed physicians' confidence in their first and second (individual) diagnoses. This section analyses doctors' confidence that their generated set of hypotheses contained the correct diagnosis and relates confidence in the set of diagnoses to their confidence in the first and second hypothesis. Overall confidence in the set was not a function of the size of the set, i.e., there was no relationship between overall confidence and the total number of diagnoses doctors generated. The amount of clinical information indicative of diseases A and B and of background information indicative of disease B all had significant effects on doctors' confidence in their set of hypotheses. The same pattern of results occurred for each case, even though to different degrees. Getting stronger clinical information indicative of A increased doctors' confidence in their set, whereas getting stronger clinical or background information indicative of B decreased doctors' overall confidence. These results suggest that confidence in the generated set of hypotheses may be a reflection of the degree of conflict doctors experience between the competing hypotheses A and B. Factors that increased the distance between the competing hypotheses (i.e., factors that increased the strength of the A-diagnosis and decreased the strength of the B-diagnosis) resulted in an increase in overall confidence in the set.

Overall confidence in the set of diagnoses was highest (mean = 8.25) when both the first and second diagnoses were of an identical type (i.e., both A- or both B-diagnoses). It was significantly lower (mean = 7.2) when they were of a different type (i.e., one A-type and the other a B-type). Further evidence for the conflict interpretation of overall confidence comes from two fact that, across physicians and cases, confidence in the set of generated hypotheses increased when the primary A-diagnosis was listed earlier ($r(250) = -0.17$, $p < 0.01$) and when the secondary B-diagnosis was listed later on the list of possible diagnoses ($r(249) = 0.12$, $p < 0.05$).

4.7. INDIVIDUAL DIFFERENCES

Hypothesis generation and confidence in generated hypotheses were also analysed for effects of individual differences in our set of diagnosticians.

4.7.1. *Effects of experience*

Weber *et al.* (1993) found that availability of a hypothesis by having diagnosed a similar case before, made doctors generate this diagnosis earlier and more numerously. Experience of physicians affected hypothesis generation solely by increasing the availability of similar cases. That is, more experienced physicians were more likely to answer the question whether they had seen a similar case before in the affirmative, and their diagnosis on that occasion affected the likelihood with which they generated different hypotheses. There was no further effect of experience when availability was included as a covariate in the analysis.

The effect of availability, i.e., of having seen a similar case before, was much weaker on physicians' confidence than on hypothesis generation. While confidence was greater when a similar case had been seen before, and especially when this case had been an A-type, this difference remained significant only for case I after controlling for the experimental information factors.

Even though there were only weak effects of the availability of a similar case, overall confidence in the set of generated hypotheses was strongly related to the amount of clinical experience of the diagnosticians as shown in Figure 6 (top panel), suggesting that these effects of experience must be mediated by factors other than availability. Overall confidence in the diagnostic set first increased with years of clinical practice, peaking at around 15 years, and then decreased again, a quadratic pattern that remained significant after partialling out the effects of the information variables and the availability of a prior case.[11] There was no interaction between experience and any of the information conditions in their effects on confidence.

Experience had a similar nonmonotonic effect on physicians' confidence in their second diagnosis (see Figure 6, bottom panel). Diagnostic confidence increased with years of clinical practice, peaking at around 17 years, and then decreased again, a quadratic pattern that remained significant after partialling out the effects of the information variables and the availability of a prior case. Experience had a similar but nonsignificant effect on physicians' confidence in their first diagnostic hypothesis.

4.7.2. Effects of gender

All three judgments of confidence were significantly lower for female than for male physicians, as shown in Figure 7.[12] There was again a dissociation in effects on hypothesis generation and confidence. Whatever factors may have given rise to the

[11] Even though physicians were randomly assigned to the information conditions of this study, without any consideration of their level of clinical experience, such assignment could have failed to achieve equality. By chance, experience may not have become distributed equally across the 16 different experimental cells, with the result that some apparent effects of experience could have resulted a confound between experience level and experimental conditions. To detect such confounds, experience (years of clinical practice) was analysed as a function of the four information factors and their interactions in an ANOVA. Only one of the 15 (main and interaction) variables, namely clinical information indicative of diagnosis B, showed significant differences in level of experience, but in a direction that would predict an effect on confidence opposite to the one observed. Thus the observed quadratic pattern was not due to any experience–information condition confound.

[12] This gender effect was partly due to a confound with information conditions, i.e., to a non-even distribution of the 16 clinical and background information conditions across male versus female physicians. An ANOVA of gender on the four information factors and their interactions showed that female doctors were more likely to have received cases with full clinical information indicative of a B-diagnosis and with reduced background information indicative of an A-diagnosis, both conditions that reduce confidence as described above. There also was a significant quadratic relationship between gender and amount of clinical experience, with the proportion of male doctors increasing over the whole range of years of clinical practice, but more steeply at the beginning. However, even after partialling out the effects of experimental information conditions and level of experience, the effect of gender on confidence remained significant for all three judgments of confidence.

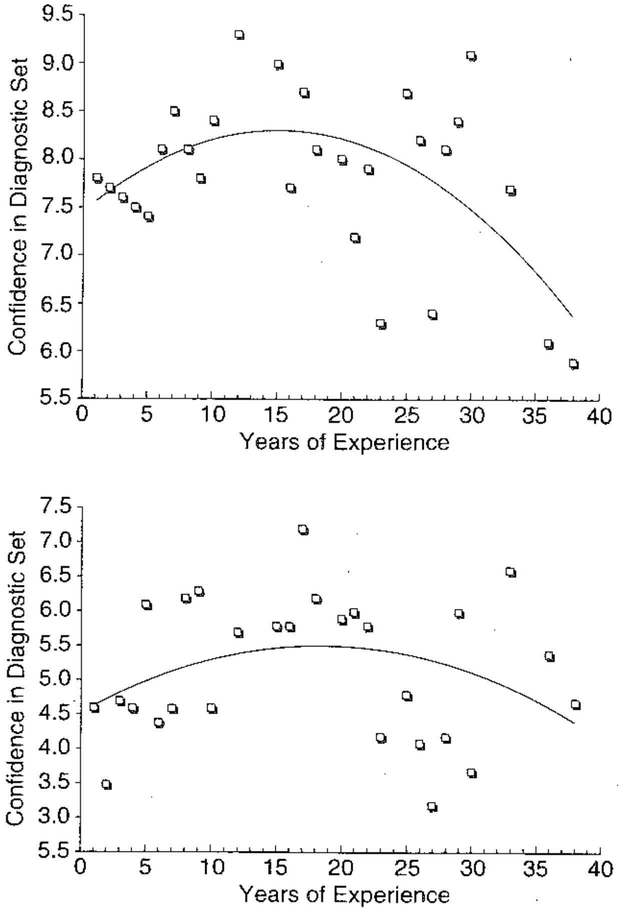

Figure 6 Relationship between physicians' experience (i.e., years of clinical practice) and overall confidence in generated set of hypotheses (top panel) and confidence in second diagnosis. Scatterpots show mean confidence across respondents and cases for each observed level of experience together with best-fitting quadratic function.

reduced confidence levels of female physicians (including the confound with the information factors) did not result in any observed differences in hypothesis generation.

One possible mechanism for this gender gap in confidence arises from the fact that female doctors were significantly more likely to mention a rival hypothesis in their justifications of their confidence in their first hypothesis than their male counterparts (29% versus 16%; $\chi^2(1) = 4.67$, $p < 0.04$). Given that, regardless of gender, there was a negative correlation between the mention of a rival hypothesis and confidence in the first diagnosis ($r(251) = -0.26$, $p < 0.0001$) and a positive correlation between confidence in first diagnosis and overall confidence in set, this provides a possible mechanism for the lower confidence ratings observed for the female doctors in our sample. It raises the question, however, of why female physicians were more likely to consider a rival hypothesis. *Post-hoc* reasons include gender differences in cognitive style, e.g., in need for cognition (Cacioppo and Petty, 1982), with individuals with a

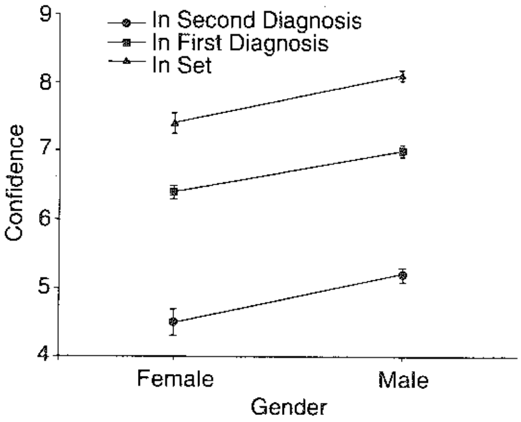

Figure 7 Mean confidence levels in the set of diagnostic hypotheses and in their first and second hypotheses, as a function of the gender of the diagnosing physician.

higher need for cognition (i.e., women) known to engage in more systematic processing (Cacioppo *et al.*, 1983). It could also be related to the more recent training of female physicians who on average were younger than male physicians. Such a confound between gender on the one hand and education, professional training, and experience on the other hand would be consistent with the results of Gigerenzer *et al.* (1991) and Lichtenstein and Fischhoff (1981) who found that male respondents showed both greater confidence and greater accuracy than female respondents. However, the effect of gender remained significant on all three confidence judgments after partialling out the effect of experience (as both a linear and a quadratic factor) from the prediction equation. Thus gender differences in confidence were not due to differences in experience, and there was no corresponding difference in hypothesis generation.

4.8. ROLE OF CONFIDENCE IN DYNAMIC DECISION MAKING

Doctors' confidence in their set of diagnoses as well as in the first two individual hypotheses was associated with differences on a variety of subsequent information preference measures. After providing justifications for their judgments of confidence and significance in their first- and second-listed diagnoses, respectively, physicians were asked, for both their first and second diagnosis, what additional specific items of information they would require before deciding upon it as their working diagnosis. Consistent with an interpretation of confidence as reflecting experienced conflict between competing hypotheses, doctors who expressed greater overall confidence in their set of hypotheses were less likely to mention a rival hypothesis in their open-ended discussion about additional information they might want before deciding on their first hypothesis as a working diagnosis.

There was also evidence consistent with attempts to establish some explanatory coherence between hypothesized diagnoses and presented symptoms. A presented piece of clinical information or the absence (or nonmention) of some clinical symptom may not fit with the diagnosis considered for adoption as a working hypothesis. Such lack of coherence could be resolved either by postulating a rival hypothesis or by further investigating the 'nonfitting' symptoms to determine the accuracy of the observation. The physicians of our study seemed to follow one strategy or the other, but not both. Some mentioned a rival hypothesis in their discussion of what additional information they might want before deciding on their second hypothesis as their working diagnosis, and some asked for additional clinical information or asked to further investigate one of the pieces of provided information. Those who mentioned a rival hypothesis were less likely to ask for additional clinical information or to ask to further investigate one of the pieces of provided information.

The differences in confidence associated with gender were accompanied by differences in subsequent information requests. Whether as a result of their lower confidence or as the result of the same factors giving rise to lower confidence (e.g., greater caution due to less experience or greater perceived need to justify their decisions), female physicians asked for more pieces of clinical or laboratory information than their male counterparts in their listing of the additional information they would require before being willing to settle on their diagnosis as a working hypothesis.

5. Discussion

5.1. ROLE AND PURPOSE OF CONFIDENCE

Confidence can be useful as an explanatory construct in a model of judgment or choice in different ways. In a weak sense, the concept of confidence can be useful by providing a framework that unifies the existence of different effects of input variables on responses. Chaiken *et al.* (1989), for example, propose a social–psychological model of information processing in attitude formation that postulates that people have a 'sufficiency threshold' for the level of confidence that they aim to achieve in their attitudes. The assumption of a sufficiency threshold for confidence allows them to explain a large number of effects of input manipulations on final attitudes and allows them to attribute these effects to a single underlying mechanism, namely the trade-off between people's wish for valid conclusions and their desire to minimize analytical effort. Another example is the Index of Consumer Sentiment, colloquially referred to as index of consumer confidence, developed by George Katona at the University of Michigan's Institute for Survey Research in the 1950s (see Converse, 1990, pp. 353–6) and used by economists in the spirit of a unifying framework for the effects of other, more objective, variables. Consisting of attitudinal questions as well as questions about buying intentions, the index was designed as a leading indicator of changes in consumer expenditure patterns. While consumer attitudes largely failed to provide predictive power above and beyond that of the economic variables shaping the attitudes – e.g., income, inflation rate, or stock prices (Tobin, 1959; Shapiro, 1972), the popular business press continues to report the level of consumer confidence as a conceptually more intuitive and convenient summary measure.

At a process level, knowledge about confidence can also be useful if confidence 'drives' people's responses and further actions. The subjective feeling of hunger, for example, is an important intervening variable in explaining food intake, even though much of the variance in meal-taking can be explained by objective variables like blood sugar level. Yet it is the extent of hunger that motivates people to search for food. Correspondingly, objective external variables may completely account for a person's choices or judgments at the level of predicting them, but at a process level individuals may actually use their level of confidence to decide on their response. Confidence then remains an important intervening construct because of the possibility that confidence levels may be affected by variables other than the normal set of inputs, just as hunger and resulting food intake may be triggered by factors other than blood sugar level (e.g., by seeing an appetizing food advertisement).

Confidence is useful in a stronger sense if no simple mapping between information input and decision output could explain the output as well as when confidence is added as a mediating construct – analogous to hidden units in a connectionist network than can account for complex interactions between input variables. In this case, some knowledge of confidence is necessary for the prediction of subsequent behaviour. Alternatively, confidence may not be necessary to predict the *outcome* of the decision, but it may carry information about the decision *process* that may not have been incorporated into the final choice but could be of interest to predict other subsequent behaviour (e.g., the acquisition of additional information or other hedging actions).

The picture of confidence in diagnostic hypotheses that emerges from the present study suggests that confidence is useful in this stronger sense of the word. Judgments of confidence were found to carry important information about the nature of the processes that gave rise to the diagnostic selection (in particular the degree of conflict between competing diagnoses) that was not available by looking at generation proportions, yet determined at least some subsequent diagnostic behaviours.

5.2. CONSEQUENCES OF CONFIDENCE

The confidence that physicians expressed in their generated diagnostic hypotheses predicted some subsequent behaviours. Physicians who expressed less confidence in a given diagnosis were more likely to mention an alternative hypothesis and asked for more additional information before being willing to settle on that diagnosis as their working hypothesis. While these requests were hypothetical (i.e., made on a paper-and-pencil test), other evidence suggests that confidence can be predictive of real-world differences in information usage. Weber and Sonka (1994) examined the relationship between judgments of confidence and farm production practices and pricing methods (assessed both from self-reports and accounting data) in a sample of Midwestern farmers. Farmers judged their confidence in their ability to give predictive weather and climate forecasts, both for specific forecasts (e.g., average maximum temperature for July predicted in March) and in general. The two confidence judgments predicted a variety of actual production and pricing decisions. Greater confidence in their own intuitive climate forecasts, for example, resulted in more frequent use of their own rather than professional forecasts, made it less likely that farmers used the futures market, and made it more likely that they used an after-harvest pricing strategy of their own design rather than one suggested by an outside consultant. While the evidence is only correlational, in conjunction with the results of this study it suggests that at least some real-world decisions and practices can be partially predicted from (and may be mediated by) confidence levels in a judgment or decision. Van Wallendael and Guignard (1992) similarly found that, in a categorization task in which people could buy information varying in cost and diagnosticity, information requests were best explained by the use of a confidence criterion adjusted for information cost.

5.3. CONFLICT, COHERENCE, AND CONFIDENCE

Our evidence of the role of diagnostic conflict in determining post-generation confidence is consistent with Zakay's (1985) conflict model of post-decision confidence in a multiattribute decision-making task. In Zakay's (1985) study, student nurses were asked for their decisions in hypothetical but typical work situations. The multiattribute decision alternatives were designed to present respondents with serious tradeoff problems (i.e., conflict) if they used a compensatory decision strategy, but less conflict if they used a noncompensatory decision strategy. Ratings of post-decisional confidence were significantly higher for choices based on noncompensatory processing that entailed less decision conflict than for choices based on a compensatory strategy with its conflictual tradeoffs. In a multiattribute decision-making task under time pressure (Böckenholt and Kroeger, 1993), post-decision confidence was positively related to the difference in attractiveness between the two choice alternatives, presumably because choices engender more conflict when the alternatives are more similar in attractiveness. These results suggest that confidence judgments in multiattribute choice and in diagnostic hypothesis generation may both express characteristics of the process leading up to the decision or judgment, in particular the subjective experience of conflict, even though conflict can be generated by different variables, depending on the task.

Just like confidence judgments for single hypotheses, confidence judgments in the set of diagnoses seemed to express diagnostic conflict. Holding two strongly competing hypotheses seemed to present interpretive 'ambiguity', in the sense that the presented set of symptoms could be related to two different and usually mutually exclusive diagnoses. Given that the diagnoses had different implications for treatment in our study and that one of them had serious clinical consequences if not followed up, the ambiguity of partial evidence for both diagnoses may have sent a signal about the difficulty of correctly diagnosing this case that lowered confidence in the generated set of hypotheses. There may be a connection between the interpretation of confidence as reflecting the perceived degree of conflict between hypotheses and Einhorn and Hogarth's (1986) model of causality, according to which a hypothesis is seen as 'causal' to the extent to which there are no rival hypotheses. Both may be manifestations of a desire for simplicity by which phenomena seem 'more explained' if only one strong hypothesis or diagnosis exists, rather than conflicting explanations of close to equal strength. Kelley's (1973) discounting model, which assumes that confidence in a given cause is decreased in the presence of a rival cause, is consistent with this interpretation.

Evidence that physicians may be more comfortable (and thus express greater confidence) with a single strong diagnostic hypothesis than with competing hypotheses comes from a study by Wolf *et al.* (1985) where physicians were asked to decide which of two diagnoses was more likely for a patient known to have two symptoms. They were provided with the probability of one diagnosis given one of the two symptoms. One additional piece of information could be requested from a total of three: the probability of the other diagnosis given the same symptom, or the probability of either one of the two diagnoses given the other symptom. Only 24% of the 89 respondents selected the information consistent with a differential diagnosis (i.e., the probability of the other diagnosis given the first symptom) for all three cases that they saw. The majority of physicians selected information about the probability of the same diagnosis given the other symptom. Of those who selected the information consistent with a differential diagnosis, 97% decided on the correct diagnosis, i.e., the one more likely according to Bayes' theorem. Those physicians who selected a nonoptimal piece of information performed only at chance (53%). Selection of the optimal piece of information is consistent with a competing-hypothesis mindset. However, the majority of physicians preferred to pursue information related to only one diagnosis, with a resulting decrement in diagnostic accuracy.

When judging confidence, physicians were sensitive primarily to the pattern of the provided information and its fit with different diagnoses, without much attention to population base rates. This result is consistent with the evidence by Griffin and Tversky (1992) that people focus on the strength of available evidence with insufficient regard for its weight when providing judgments of confidence.

Pennington and Hastie (1986) found that coherence affects confidence in the context of jury decision making: the confidence of jurors in their verdict was a function of the difference in coherence between the prosecution's story and the story presented by the defense. Recent work on expert systems in medicine (Evans and Gadd, 1989) has started to acknowledge the evaluation of the coherence of information as an important component of diagnostic expertise. The present results suggest that physicians' judgments of their diagnostic confidence would provide important sources of information for the development of expert systems concerned with the evaluation of explanatory coherence.

5.4. EXPERIENCE AND CONFIDENCE

Clinical experience had a nonmonotonic effect on confidence for which only *post hoc* explanations can be offered. The initial increase and eventual decrease in confidence with years of clinical practice may be caused by different and possibly independent factors. Increases in confidence with experience could be due to growing expertise with symptom – diagnosis relationships, resulting in greater ability to make more coherent sense out of symptom configurations. Weber and Sonka (1994), for example, found such a positive effect of experience on the confidence that farmers had in their intuitive climate forecasts over the whole range of experience (from 4 to 60 years).

Decreases in confidence with clinical experience could be due to different processes that may occur in addition to (and superimposed on) this positive effect. If these negative processes grew at a faster rate than the positive processes or started at a later time, they would eventually lead to a reversal in the direction of the effect. Decreases in confidence could be the result of increasing concern about the possibility of diseases with serious clinical consequences that are not frequently observed but ignored at considerable peril. Because such diseases tend to have low population base rates, learning about them and their consequences might take many years of clinical practice or even some personal experience with severe illnesses that comes with increasing age, as shown by Weber *et al.* (1993). Greater awareness of such competing hypotheses would reduce confidence according to the conflict model proposed in this paper.

Alternatively, physicians may realize after having been out of medical school for a while, that their knowledge has become outdated (Evans *et al.*, 1984), leading to a decrease in their confidence in judgment. Also, due to the cross-sectional nature of our sample of physicians with different levels of experience, older doctors may simply be less skillful or less schooled, and therefore less confident. Berwick *et al.* (1981), for example, administered a test of quantitative clinical reasoning to practicing physicians and found a highly significant negative correlation between test performance and years of clinical experience, which ranged from 2 to 48 years. Among other explanations, they suggest that this may be the result of a loss of skill over time or of increasing stringency in entrance or certification requirements. If physicians were aware of either of these factors, their awareness may result in the decrease in confidence observed in this study.

5.5. IMPLICATIONS FOR REDUCTION OF OVERCONFIDENCE

When subjective confidence in single judgments can be compared to long-run objective accuracy, confidence has been found to be poorly calibrated in a variety of different contexts and domains (Adams and Adams, 1961; Lichtenstein and Fischhoff, 1977; Yates, 1990). In these studies where people answer knowledge questions and indicate confidence in their answer as a percentage figure or by placing a confidence interval around their point estimate, answers are less often accurate than expected based on the percentage indicated by the confidence judgment or interval.

Remedial techniques that have been found to reduce (over)confidence to more accurate levels (Koriat *et al.*, 1980; Hoch, 1985; Arkes *et al.*, 1987) generally instigate more extensive processing of all available choice alternatives, for example, by requiring respondents to generate reasons why their selected answer as well as alternative answers may either be correct or false or by getting people to argue about their answers in a group discussion. The results of our study suggest that it may not be the extent of the processing *per se* that reduces confidence levels, but the experience of conflict between alternatives that may or may not result from more extensive processing or group discussion. This interpretation makes the prediction that overconfidence will be reduced only to the extent that further processing of (initially rejected) alternatives reveals them as more plausible than previously thought (i.e., reveals previously concealed conflict between competing responses). It is consistent with the results by Hoch (1985) who found that having MBAs generate reasons against the accuracy of their predictions was effective in reducing their overconfidence for low to moderate base rate events, but not for high base rate events. If confidence reflects the subjective experience of the process by which an answer or estimate is generated, confidence should not only be greater for high base rate than for low base rate events because the former are easier to generate, but should also be less affected by counter arguments because those will induce less conflict for high base rate events, as found in our study. Also consistent are the results by Sniezek *et al.* (1990) that confidence in an answer is significantly lower when the judge selected that answer herself out of a set of possible answers than when it was selected by the experimenter. Presumably, the act of choosing the alternative for which confidence is judged exposes people more to possible conflict between choice alternatives.

5.6. CONFIDENCE AS THE 'MEMORY' OF THE DYNAMIC DECISION PROCESS

The results of our study are consistent with an interpretation of confidence as an expression of characteristics of the process that gave rise to the judgment or the decision, in this case to the generated diagnosis. Just as for the simple psychophysical judgments discussed in the introduction, confidence in diagnostic decisions seems to reflect the ease or difficulty experienced over the course of the decision process. Confidence in a single judgment was affected by the degree of coherence or conflict between the available evidence and the given diagnostic judgment or decision. Confidence in a set of hypotheses was affected by the degree of conflict between competing members of the set. Goldstein and Weber (1995) found some direct evidence of a relationship between decisional confidence and decisional effort (i.e., a negative correlation; $r(288) = -0.50$, $p < 0.0001$) in a task where people rated potential marriage partners and provided separate assessments of their confidence in their final judgment and the effortfulness of the decision.

Since confidence judgments and generation frequencies of the diagnoses were affected in qualitatively different ways by the informational factors manipulated in our study, as well as by individual-difference characteristics of the physicians, judgments of confidence add information to the understanding of the diagnostic process above and beyond simply knowing which diagnosis a physician selected. Confidence judgments and generation frequencies differed to the point of very different rank-orders across the different information conditions, especially for the low-base rate B-diagnoses. Both measures differed from the predictions of a Bayesian model of the probability of being correct. Generation proportions failed to reflect the degree of conflict due to incoherence between inconsistent symptom information and the selected diagnosis. Doctors generated a given diagnosis with equal frequency regardless of the amount of information inconsistent with that diagnosis, i.e., regardless of the likelihood of competing (but weaker) hypotheses. The degree of conflict preceding this final resolution, however, found expression in physicians' judgments of their confidence in their diagnosis.

Just as for simple psychophysical tasks (e.g., visual or other discrimination tasks) where researchers have found little evidence of a relationship between confidence judgments and overall accuracy (Garrett, 1922; Johnson, 1939; Pierrel and Murray, 1963; see Link, 1992, Chapter 5, and Vickers, 1979, Chapter 6, for an overview), the present study showed sizable deviations between confidence judgments and Bayesian predictions of the probability of being correct, especially for hypotheses with low population base rates. The image of the strongest response competitor winning out in the struggle for generation, with competing voices getting a hearing primarily and almost exclusively in post-generation judgments of confidence is consistent with characterizations of discrimination or other choices in terms of random walk models (e.g., Estes, 1960; Link and Heath, 1975), where confidence describes the length of quality of the 'walk,' i.e., serves as a measure of the 'balance of evidence' (Vickers, 1979, p. 176).

The precise nature in which confidence serves as a 'memory' or evaluation of the episode of processing in more complex judgment and decision tasks deserves further investigation, along the lines of a recent study by Zakay and Tsal (1993). A new (or rather renewed) interpretation of confidence as a reflection of characteristics of the decision process may also help explain some of the ways in which the magnitude of confidence judgments seems to depend on the nature of the task. Different tasks result in different processes, with different factors affecting the ease of the process leading to final judgment or decision. If confidence reflects an evaluation of this process, 'some kind of differentiation and discrimination of internal cues – feelings of doubt' (Adams and Adams, 1961, p. 43), then task characteristics clearly should affect confidence.

Acknowledgements

This research was supported by Biomedical Research Support Grant RR 7030 from the National Institute of Health and by a grant from the University of Illinois Research Board to the first three authors. We thank Paul Heckerling and Barbara Shapiro for sharing their medical expertise, and Sema Barlas, Bill Goldstein, Barbara Mellers, and Lori Van Wallendael for helpful comments and suggestions.

References

Adams, J.K. and Adams, P.A. (1961) Realism of confidence judgments, *Psychological Review*, **68**, 33–45.

Arkes, H.R., Christensen, C., Lai, C. and Blumer, C. (1987) Two methods of reducing overconfidence, *Organizational Behavior and Human Decision Processes*, **39**, 133–44.

Bar-Hillel, M. (1980) The base rate fallacy in probability judgments, *Acta Psychologica*, **44**, 211–33.

Barrows, H.S., Norman, G.R., Neufeld, V.R. and Feightner, J.W. (1982) The clinical reasoning of randomly selected physicians in general medicine practice, *Clinical and Investigative Medicine*, **5**, 49–55.

Beach, B.H. (1975) Expert judgment about uncertainty: Bayesian decision making in realistic settings, *Organizational Behavior and Human Performance*, **14**, 10–59.

Berwick, D.M., Fineberg, H.V. and Weinstein, M.C. (1981) When doctors meet numbers, *American Journal of Medicine*, **71**, 991–98.

Böckenholt, U. and Kroeger, K. (1993) The effect of time pressure in multiattribute binary choice tasks, in O. Svenson and A.J. Maule (eds) *Time pressure and stress in human judgment and decision making*.

Böckenholt, U., Albert, D., Aschenbrenner, K.M. and Schmalhofer, F. (1991) The effect of attractiveness, dominance, and attribute differences on information acquisition in multiattribute binary choice, *Organizational Behavior and Human Decision Processes*, **49**, 258–81.

Cacioppo, J.T. and Petty, R.E. (1982) The need for cognition, *Journal of Personality and Social Psychology*, **42**, 116–31.

Cacioppo, J.T., Petty, R.E. and Morris, K.J. (1983) Effect of need for cognition on message evaluation, recall, and persuasion, *Journal of Personality and Social Psychology*, **45**, 805–18.

Chaiken, S., Liberman, A. and Eagly, A.H. (1989) Heuristic and systematic information processing within and beyond the persuasion context, in J.S. Uleman and J.A. Bargh (eds) *Unintended Thought*, New York: Guilford Press.

Clarke, F.R. (1960) Confidence ratings, second-choice responses, and confusion matrices in intelligibility tests, *Journal of the Acoustical Society of America*, **32**, 35–46.

Converse, J.M. (1990) *Survey Research in the United States: Roots and Emergence 1890–1960.* Berkeley, CA: University of California Press.

Dawes, R.M. and Mulford, M. (1996) The false consensus effect and overconfidence: flaws in judgment or flaws in how we study judgment? *Organizational Behavior and Human Decision Processes,* **65**, 201–11.

Einhorn, H.J. and Hogarth, R.M. (1986) Judging probabily cause, *Psychological Bulletin,* **99**, 3–19.

Elstein, A.S., Shulman, L.S. and Sprafka, S.A. (1978) *Medical Problem-Solving: An Analysis of Clinical Reasoning,* Cambridge, MA: Harvard Press.

Estes, W.K. (1960) A random walk model for choice behavior, in K.J. Arrow, S. Karlin and P. Suppes (eds) *Mathematical Methods in the Social Sciences 1959,* Stanford, CA: Stanford University Press.

Evans, D.A. and Gadd, C.S. (1989) Managing coherence and context in medical problem-solving discourse, in D.A. Evans and V.L. Patel (eds) *Cognitive Science in Medicine,* Cambridge, MA: MIT Press.

Evans, C.E., Haynes, R.B., Gilbert, J.R., Taylor, D.W., Sackett, D.L. and Johnston, M. (1984) Educational package on hypertension for primary care physicians, *Canadian Medical Association Journal,* **130**, 719–22.

Fischhoff, B., Slovic, P. and Lichtenstein, S. (1977) Knowing with certainty: The appropriateness of extreme confidence, *Journal of Experimental Psychology: Human Perception and Performance,* **3**, 552–64.

Garrett, H.E. (1922) A study of the relation of accuracy to speed, *Archives of Psychology,* **56**, 1–105.

Gigerenzer, G., Hoffrage, U. and Kleinboelting, H. (1991) Probabilistic mental models: a Brunswikian theory of confidence, *Psychological Review,* **98**, 506–28.

Goldberg, L.R. (1959) The effectiveness of clinicians' judgments, *Journal of Consulting Psychology,* **23**, 25–33.

Goldstein, W.M. and Weber, E.U. (1995) Content and discontent: Indications and implications of domain specificity in preferential decision making, in J.R. Busemeyer, R. Hastie and D.L. Medin (eds) *Decision Making from a Cognitive Perspective. The Psychology of Learning and Motivation,* Vol. 32 (pp. 83–136), New York: Academic Press.

Griffin, D. and Tversky, A. (1992) The weighing of evidence and the determinants of confidence, *Cognitive Psychology,* **24**, 411–35.

Heath, F. and Tversky, A. (1991) Preference and belief: ambiguity and competence in choice under uncertainty, *Journal of Risk and Uncertainty,* **4**, 5–28.

Henmon, V.A.C. (1911) The relation of the time of a judgement to its accuracy, *Psychological Review,* **18**, 186–201.

Hoch, S.J. (1985) Counterfactual reasoning and accuracy in predicting personal events, *Journal of Experimental Psychology: Learning, Memory, and Cognition,* **11**, 719–31.

Janis, I.L. and Mann, L.A. (1968) A conflict theory approach to attitude change and decision making, in A. Greenwald, T. Brock and K. Ostram (eds) *Psychological Foundations of Attitudes,* New York: Academic Press.

Johnson, D.M. (1939) Confidence and speed in the two-category judgment, *Archives of Psychology,* **34**, 1–53.

Kelley, H.H. (1973) The process of causal attribution, *American Psychologist,* **28**, 107–28.

Koriat, A., Lichtenstein, S. and Fischhoff, B. (1980) Reasons for confidence, *Journal of Experimental Psychology: Human Learning and Memory,* **6**, 107–18.

Lichtenstein, S. and Fischhoff, B. (1977) Do those who know more also know more about how much they know? The calibration of probability judgments, *Organizational Behavior and Human Performance,* **20**, 159–83.

Lichtenstein, S. and Fischhoff, B. (1981) Cited in text.

Link, S.W. (1992) *The Wave Theory of Difference and Similarity*, Hillsdale, NJ: Lawrence Erlbaum.

Link, S.W. and Heath, R.A. (1975) A sequential theory of psychological discrimination, *Psychometrika*, 1, 77–105.

Oskamp, S. (1962) The relationship of clinical experience and training methods to several criteria of clinical prediction, *Psychological Monographs: General and Applied*, **76**, 1–27.

Oskamp, S. (1965) Overconfidence in case-study judgments, *Journal of Consulting Psychology*, **29**, 261–65.

Paese, P.W. and Sniezek, J.A. (1991) Influence on the appropriateness of confidence in judgment: Practice, effort, information, and decision making, *Organizational Behavior and Human Decision Processes*, **48**, 100–30.

Peirce, C.S. and Jastrow, J. (1884) On small differences in sensation, *Annals of the National Academy of Science*, **3**, 73–83.

Pennington, N. and Hastie, R. (1986) Evidence evaluation in complex decision making, *Journal of Personality and Social Psychology*, **51**, 242–58.

Peterson, D.K. and Pitz, G.F. (1988) Confidence, uncertainty, and the use of information, *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **14**, 85–92.

Pierrel, R. and Murray, C.S. (1963) Some relationships between comparative judgment confidence and decision time in weight lifting, *American Journal of Psychology*, **76**, 28–38.

Robins, R.W. and Craik, K.H. (1993) Is there a citation bias in the judgement and decision making literature? *Organizational Behavior and Human Decision Processes*, **54**, 225–44.

Schwarz, N., Bless, H., Strack, F., Klumpp, G., Rittenauer-Schatka, H. and Simons, A. (1991) Ease of retrieval as information: another look at the availability heuristic, *Journal of Personality and Social Psychology*, **61**, 195–202.

Shapiro, H.T. (1972) The index of consumer sentiment and economic forecasting: a reappraisal, in B. Strumpel, J.N. Morgan and E. Zahn (eds) *Human Behavior in Economic Affairs: Essays in Honor of George Katona*, San Francisco, CA: Jossey-Bass Publishers.

Sniezek, J.A. and Henry, R.A. (1989) Accuracy and confidence in group judgment, *Organizational Behavior and Human Decision Processes*, **43**, 1–28.

Sniezek, J.A., Paese, P.W. and Switzer, F.S. (1990) The effect of choosing on confidence in choice, *Organizational Behavior and Human Decision Processes*, **46**, 264–82.

Soll, J.B. (1996) Determinants of overconfidence and miscalibration: the role of random error and ecological structure, *Organizational Behavior and Human Decision Processes*, **65**, 117–37.

Tobin, J. (1959) On the predictive value of consumer intentions and attitudes, *The Review of Economics and Statistics*, **41**, 1–11.

Van Wallendael, L.R. and Guignard, Y. (1992) Diagnosticity, confidence, and the need for information, *Journal of Behavioural Decision Making*, **5**, 25–37.

Vickers, D. (1979) *Decision Processes in Visual Perception*, Chapter 6 (pp. 171–200), New York: Academic Press.

Weber, E.U., Böckenholt, U., Hilton, D.J. and Wallace, B. (1993) Determinants of hypothesis generation: effects of information, baserates, and experience, *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **19**, 1–14.

Weber, E.U. and Sonka, S. (1994) Production and pricing decisions in cash-crop farming: Effects of decision traits and climate change expectations, in B.H. Jacobsen, D.E. Pedersen, J. Christensen and S. Rasmussen (eds) *Farmers' Decision Making: A Descriptive Approach*, pp. 203–18), Copenhagen, Denmark: European Association of Agricultural Economists.

Wolf, F.M., Gruppen, L.D. and 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**, 2858–62.

Yates, J.F. (1990) *Judgment and Decision Making,* Chapter 4; Accuracy (pp. 75–111), Englewood Cliffs, NJ: Prentice Hall.

Zakay, D. (1985) Post-decisional confidence and conflict experienced in a choice process, *Acta Psychologica*, **58**, 75–80.

Zakay, D. and Tsal, Y. (1993) The impact of using forced decision-making strategies on post-decisional confidence, *Journal of Behavioral Decision Making*, **6**, 53–68.