11

Expert Judgement of Probability and Risk

George Wright, Fergus Bolger and Gene Rowe

This chapter reviews extant research on the quality of expert judgement of probability and risk. Both types of judgement are of great relevance to the concerns of this book, because expert judgement under uncertainty is a key component in the making of decisions with scant information.

In the first section, we focus on the conditions under which experts have produced high quality assessments of probability – in terms of both the coherence and validity of assessed probabilities. We conclude that invalid probability judgements are common when features of the judgement domain are unfavourable and "learnability" is low – by which we mean that there is a lack of relevant background data and/or outcome feedback upon which to base and revise domain models. We identify judgement situations when such conditions are likely to be prevalent. We also show that use of ill-matched probability elicitation methods are a common operational flaw. We propose a task taxonomy that can be used to identify when expert probability judgements are likely to be well-made.

Next, we consider whether expert judgements of risk – by which we mean judgements of the likelihood of occurrence of hazardous events – are of higher quality than those made by members of the general public. We evaluate the nine empirical studies that have been conducted on expert versus lay judgements of risk, and find that there is little empirical evidence for the propositions (1) that experts judge risk differently from members of the public or (2) that experts are more veridical in their risk assessments. We discuss the nature of expertise and consider whether the commonsense assumption of the superiority of expert risk assessors in making magnitude judgements is, in fact, sensible. We end with a discussion of future research directions.

T. M. Williams et al. (eds.), *Making Essential Choices with Scant Information*

Expert judgement of probability

Expert judgement about uncertainty can be expressed and elicited in many ways – for example, as verbal probability expressions, numerical probabilities, odds, certainty factors, fuzzy sets and second order probabilities. The literature on behavioural decision-making has focused almost entirely on subjective probability because of the philosophical attractiveness of the axiom base which leads, logically, to expected utility as a choice principle. Expected utility underpins decision analysis (Goodwin and Wright, 2004), while subjective probabilities are used in forecasting techniques such as cross-impact analysis (Dalkey, 1972; Wright and Ayton, 1987). In expert system approaches to decision support, subjective probabilities are also a widely used way of representing the degrees of belief of experts (Kanal and Lemmer, 1986; Lemmer and Kanal, 1988; Shafer, 1987; Tonn et al., 1992). However, what do we know about the quality of *expert* probability judgement? If the probability assessments elicited from domain experts are poor in some respect, then the decisions of any decision-aiding technique or system which uses these probabilities will also be poor. Alternatively, if experts are capable of providing high quality probability assessments then poorly designed elicitation techniques may degrade this ability. Ensuring the quality of expert probability judgement is, therefore, of major practical importance to the implementation of decision support.

Two of the most commonly used approaches to the assessment of the quality of probability judgements are *coherence* and *calibration*. By coherence, we mean the extent to which a probability assessor's forecasts conform to the axioms of probability theory. For instance, according to one axiom of probability theory – additivity – the component probabilities of a set of mutually exclusive and exhaustive events should add up to one. Another axiom is the intersection law, which states that the probability of event A and event B both happening is the product of the probability of event A multiplied by the probability of event B, given that event A has happened. More formally, $p(A \text{ and } B) = p(A)P(B|A)$. Other probability laws determine other relationships between such "complex" probabilities and their "simple" components of marginal and conditional probabilities. In studies of coherence, probability judgements are assessed in terms of the extent to which they conform to axioms, which basically means that the judgements are consistent or reliable (see Goodwin and Wright, 2004; Yates, 1990). In studies of calibration, probability judgements are validated against an external objective standard. For perfect calibration, a set of events all allocated a 0.XX probability of occurrence by a forecaster should, in actual fact, occur on XX% of occasions. For example, if we examined all the days where a perfectly calibrated weather forecaster assessed the probability of rain to be 0.7, then we should find the true proportion of rainy days also to be 0.7 (see Lichtenstein et al., 1982;

Yates, 1990 for reviews). Calibration is, therefore, a measure of the validity of a set of subjective probability assessments.

It has been argued that coherence and calibration are logically interrelated in the same way that reliability and validity are related in measurement theory (Wallsten and Budescu, 1983; Wright and Ayton, 1987). Reliability is usually regarded as a necessary but not sufficient prerequisite for validity. In terms of analogy, valid judgements must be reliable, but reliable judgements are not necessarily valid (Carmines and Zeller, 1979). For example, a metre rule which is actually 99 cm long will give reliable but non-valid measurements. However, an elastic rule which changes length from time to time is both unreliable and, generally, invalid. Thus, a coherent judge has the potential to be well-calibrated but an incoherent judge is, logically, incapable of being systematically well-calibrated.

Research on calibration and coherence

Research on the quality of calibration performance of experts' probability assessments – usually with respect to forecasting performance – has been found, in several instances, to be very good; for example, Kabus (1976) (financial interest rates); Hoerl and Fallin (1974) (horse racing); Keren (1987) (the card game, Bridge); and, as we shall discuss in detail later, most strikingly in weather forecasting – Murphy and Brown (1985). Conversely, in several instances poor calibration has been found e.g. Oskamp (1965) (clinical psychologists); Wallace (1923) (maize judges). One explanation for those instances, where poor expert calibration has been found, might be lack of coherence (recall the argument relating reliability and validity).

A number of studies have demonstrated incoherence in *students'* probability judgement using paper-and-pencil tasks (e.g. Kahneman et al., 1982; Yates, 1990), although we have located only a few studies of (in)coherence in experts. For example, Eddy (1982) reviewed medical literature on the diagnosis of breast cancer from X-rays and found that physicians misunderstood the relationship between marginal probabilities (e.g. the probability of cancer, the probability of positive test, etc.) and conditional probabilities (e.g. the probability of a positive test given cancer). However, Schafer et al. (1977) found that, in two out of three tests, people who self-rated as more knowledgeable than their peers in soccer and statistics were slightly more coherent than those who rated themselves as less knowledgeable. In a study of coherence of self-rated experts in snooker, we found self-rated experts more coherent in their judgements of the probability of the union of two events than self-rated novices (Wright et al., 1994). Yet again, in a study of professional restaurant managers, Dube-Rioux and Russo (1988) found failure to conform to the additivity axiom with respect to the disjunction and conjunction of probable causes of restaurant failure. Finally, DuCharme and Peterson (1968) and Youseff and Peterson (1973) found that the failure to

revise probabilities in the light of new information required by Bayes' theorem was less marked in real-world tasks than had been previously found in the laboratory (e.g. Phillips and Edwards, 1966; Phillips et al., 1966).

Thus, as with calibration, the picture with respect to coherence is unclear. Experts sometimes demonstrate incoherence of the sort found in naive judges, but there is a suggestion that specific knowledge/expertise may, in some cases, lead to a reduction in the extent of this incoherence. Given this tension in the literature, it remains to be demonstrated that it is incoherence which is responsible for poor calibration. For example, it might plausibly be argued that incoherence and miscalibration are both *symptoms* of poor probabilistic judgement. In fact, there are a number of psychological explanations for poorly calibrated probability judgement (see, e.g. McClelland and Bolger, 1994). We shall examine some of these other possible causes of poor calibration performance in probabilistic judgement shortly. However, it should be noted that there is a sizeable literature that shows random error in probability judgements can explain much of the observed miscalibration, although not all (see Ayton and McClelland, 1997 for a review).

Decomposition-recomposition

Most procedures for producing coherent probability judgements involve a technique called decomposition-recomposition. For example, consider the abstract example of assessing the probability of drawing two consecutive aces from a pack of 52 cards. This probability (without replacement) is

$$p(2 \text{ aces}) = 4/52 * 3/51 = 12/2652 = 0.00452.$$

Intuitively, it would seem that most people could, with a little thought, accurately assess the probability of drawing an ace on the first draw and the probability of a subsequent second ace. It also seems intuitively reasonable that most people could not make an accurate mental computation of the probability of the intersection of the two events. On the basis of the observation that most incoherence is manifest when judges attempt to revise or combine probabilities, it has been assumed that it is difficulties in computation (not probabilistic estimation *per se*) which lead to errors. Thus, it has been proposed that judgement problems should be broken up (decomposed) into small elements for which judges supply probability estimates. Probability assessments are then reproduced by mechanistically combining the individual component judgements on the basis of the laws of probability theory (a process known as recomposition).

Decomposition-recomposition has been shown to produce more coherent complex probabilities than those assessed by judges holistically (e.g. Edwards et al., 1968), but are the probabilities more valid, for example, in terms of calibration? Unfortunately, there is little empirical evidence available to answer

this question. Wright et al. (1988) found that calibration of unions, intersection and disjunctions of events which were mechanically recomposed from marginal and conditional assessments were no better calibrated than holistically judged forecast probabilities for the same events. Further, no significant correlation was found between the coherence and the subsequent calibration of the assessed probabilities. Wright et al.'s (1994) study also found no relationship between an assessor's degree of incoherence and his or her degree of (mis)calibration, thereby also suggesting that *increasing* a person's degree of coherence will not necessarily increase that person's calibration. Why should this be the case?

Greater reliability does not necessarily imply greater validity, although greater validity does imply greater reliability. Thus, poor calibration may be due to factors other than poor coherence. Suggestions include memory problems, use of heuristics, insensitivity to task difficulty and confirmation biases (see e.g. Compte and Postlewaite, 2004; Dougherty et al., 1999; Ferrell and McGoey, 1980; Griffin and Tversky, 1992; Tversky and Koehler, 1994). It, therefore, follows that the success of the decomposition-recomposition approach lies in the assumption that the "simple" probabilities elicited for recomposition (often marginal and conditional probabilities, as in the cross-impact technology) are themselves free of bias. If systematic error exists in these simple assessments then recomposition will magnify this bias.

Some influences on the assessed quality of judgement

Good calibration has been demonstrated in a number of instances but most notably in weather forecasting. We (Bolger and Wright, 1994) proposed that a skilled judge *can* give valid probability estimates *if* the task, elicitation and assessment procedures are amenable. Specifically, we proposed that invalid probability judgements arise when:

1. Valid probability judgement cannot easily be learned. This may be due to such influences as the amount and complexity of information, the degree to which events to be judged are related to an underlying domain model (e.g. likelihood of rain to a model of the weather) and (perhaps most importantly) the lack of outcome feedback in the task domain upon which to base and revise judgement. We term these task influences *learnability*.
2. The judge is unskilled due to lack of knowledge about the task domain and/or probability laws.
3. Probability estimates are elicited in a manner which makes them unrepresentative of the judges' true feelings of subjective probability; for example, by asking him or her to respond in an unfamiliar metric (such as odds to a non-betting person).

One instance where judgemental probability forecasts are routinely generated is weather forecasting. The official forecasts issued by the National Weather Service in the United States are subjective probability forecasts. Murphy and Brown (1985) have evaluated these subjective forecasts and found that, for certain categories of weather, they were more accurate than the available objective statistical techniques. In this case, the experimental forecasters have a very large amount of information available, including the output from statistical techniques (cf factor 2). They also receive detailed feedback (cf factor 1) and have the opportunity to gain experience of making forecasts under a wide range of meteorological conditions. Furthermore, they have considerable practice in quantifying their internal state of uncertainty (cf factor 3). These circumstances may well be ideal for the relatively successful application of judgemental compared to purely quantitative forecasting.

In our view, performance-demonstrated experience in probability judgements is underpinned by practice and regular performance feedback. However, as Einhorn and Hogarth (1978) have argued, most judgements are made without the benefit of accurate feedback. They identified three main problems:

1. The lack of search for and use of disconfirming evidence.
2. The use of unaided memory for coding, sorting and retrieving outcome information.
3. When people take an action based on a forecast in order to facilitate or avoid possible futures, they can often only observe feedback associated with the action taken, and not the action not taken.

To illustrate (3), Einhorn (1980) gives the following example:

> Imagine that you are a waiter in a busy restaurant and because you cannot give good service to all the people at your station, you make a judgement regarding which people will leave good or poor tips. You then give good or bad service depending on your judgement. If the quality of service, in itself, has an effect on the size of the tip, outcome feedback will "confirm the predictions" ("They looked cheap and left no tip – just as I thought"). The extent of such self-fulfilling prophecies is much greater than we think and represents a considerable obstacle to learning from outcome feedback.

This third feedback problem is, of course, immaterial in contexts such as weather forecasting where actions cannot be taken to increase or reduce the likelihood of the forecast event. Unconfounded feedback in such circumstances is likely to prove more useful for the improvement of forecasting ability.

Wagenaar and Keren (1986) further point out that feedback must be attended to in order for it to be of any use in improving judgement. They show that blackjack dealers were no better calibrated than "lay" people for judgements about the frequency of occurrence of certain types of hand, despite being exposed to thousands of examples each working week. They suggest that this finding is due to lack of motivation on the part of blackjack dealers to attend to the outcome feedback available to them. It is now well known that attention to the target stimuli is a necessary condition for learning about the frequency of occurrence of those stimuli.

Murphy and Brown (1985) have argued that the presence of actual or potential users of judgemental weather forecasts provides the forecasters with a strong motivation for conducting the forecasting process in an efficient and more effective manner. Moreover, feedback from users of forecasts frequently contains information regarding possible improvements. The use of judgement in real-world forecasting thus contrasts strongly with the study of judgement in the psychological laboratory, where calibration feedback and incentives for good performance have seldom been utilized.

The metric in which probability responses are elicited can take a number of different forms (e.g. percentages, point probability estimates, odds, relative frequency, etc.). Depending on which metric is used, the judge's task of turning subjective feelings of uncertainty into measurable/usable numeric estimates can be either helped or hindered. For example, Wright et al. (1988) found that for a short-term forecasting task of impersonal events (e.g. "will the pound fall below one dollar in the next 2 months?"), 29 out 36 students were better calibrated on point probabilities than odds. This experiment, therefore, gives some empirical support for the view that point probability estimates, not odds, should be elicited from untrained forecasters.

So far in this discussion of elicitation and assessment effects, we have not differentiated between the sorts of probability estimates that are being elicited (e.g. marginal or conditional? simple or complex? intersections or disjunctions?) because calibration studies have not tended to differentiate either. However, it seems to us that an important research question is the extent to which it is more natural to make some sort of probability assessments than others.

Decomposition implies that "simple" probabilities, such as marginals and conditionals, are easier for judges to assess than "complex" probabilities, such as intersections, disjunctions and unions. However, our earlier discussion of decomposition-recomposition found little evidence of a distinction between simple and complex assessed probabilities in terms of calibration performance. One possible reason for this rather surprising finding is that the problem decomposition used may not have been appropriate for the judges. In other words, the judges may have been *framing* the problem differently to the experimenter so that the decomposition did not result in

easier-to-assess probabilities. Of course, a characteristic of expertise is surely that experts can discriminate questions of a type they can answer from those they cannot (see Shanteau, 1992).

Section conclusion

In order to reduce potential invalidity due to features of the task domain, we propose that a thorough task analysis should be performed before probabilities are elicited. Questions to be addressed include the following: has the judge had the opportunity to attain good probability through experience of feedback? How long is the feedback loop? Is the feedback loop sensitive to treatment effects? To what extent are items/events related? and is it possible to validate judgements against some external standard? If the conclusion of the task analysis is that the conditions for learnability are not present in the task domain, and/or there are no available objective criteria for validating probabilistic judgement (perhaps because the events to be forecast are in the far future), then the only strategy is to ensure judgements are coherent. As we have argued, in *the absence of validity measures* the decision analyst can *only* ensure such coherence.

One practical step that can be taken to ensure that the elicitation process influences the validity of probabilistic judgement as little as possible is to encourage judges to decompose the problem in their *own* preferred way, using their *own* preferred response. In some forecasting situations, it may be that the assessor will feel happier assessing marginal and conditional probabilities than compound probabilities (cf the playing card example, earlier). However, as we have argued, in other situations the forecaster may feel more comfortable assessing compound probabilities directly. Overall, perhaps the most flexible approach to debiasing subjective probability forecasts would be to adopt what Keren (1990) calls *structure-modifying* techniques, where the user is forced/encouraged to understand the internal logic of a particular debiasing technique rather than follow a procedure blindly.

Expert judgement of risk

In a pioneering paper, Lichtenstein et al. (1978) investigated how well people (students and convenience samples from the lay population) could estimate the frequency of the lethal events that they may encounter in life.

In their study, Lichtenstein et al. (1978) found that although their subjects exhibited some competence in judging such frequencies – frequency estimates increased with increases in true frequency – the overall accuracy of both (1) paired comparisons of the relative frequency of lethal events and (2) direct estimates of frequencies of individual events were poor. In a comment on the Lichtenstein et al. study, Shanteau (1978) argued that if respondents had had more experience with the lethal events the validity of the required estimates may have shown improvement. He concluded that "It

might also be of some value to investigate judgement of lethal events, using subjects who have direct knowledge and exposure to such events (such as life insurance analysts)" (1978: p. 581).

Since the 1978 paper, research on risk judgements has led to the generally accepted conclusion that expert judgements are, indeed, more veridical than those of the general public (e.g. Slovic, 1987, 1999). One basis for this argument is the work by Slovic et al. (1985). In this study, the authors utilized samples of the US League of Women Voters, university students, members of the US Active Club (an organization of business and professional people devoted to community services activities) and a group of professional "experts". Perceptions of risk were measured by asking participants to order the 30 hazards from least to most risky (in terms of the "risk of dying (across US Society as a whole) as a consequence of this activity or technology") (1985: p. 116). Participants were told to assign a numerical value of 10 to the least risky item and to make other ratings relative to this value. Since these instructions called for a risk assessment, rather than a (relative) frequency estimate (cf Lichtenstein et al., 1978), the avenue was open – for both experts and nonexperts – for qualitative risk attributes, such as the voluntary nature or controllability of the risk, to enter into these global risk judgements.

Slovic et al. (1985) concluded that the judgement of their experts differed substantially from non-expert judgement primarily because the experts employed a much greater range of values to discriminate among the various hazards that they were asked to assess, which included motor vehicles, smoking, alcoholic beverages, hand guns, surgery, X-rays and nuclear power. Additionally, Slovic et al. (1985) concluded that their obtained expert-lay differences were "because most experts equate risk with something akin to yearly fatalities, whereas lay people do not" (1985: p. 95). This conclusion is founded on the fact that the obtained correlations between perceived risk and the annual frequencies of death were 0.62, 0.50 and 0.56 for the League of Women Voters, students and Active Club samples, respectively. The correlation of 0.92 obtained within the expert sample is significantly higher than those obtained within each of the lay samples. However, Slovic et al. (1985) also found that both the lay and expert groupings viewed the hazards similarly on qualitative characteristics such as voluntariness of risk, control over risk and severity of consequences – when asked *directly* to do so (see Rowe and Wright, 2001, for a full discussion). It would seem that when asked for a "risk" estimate, Slovic et al.'s experts viewed this as a magnitude estimation task rather than a qualitative evaluation task. Additionally, an artificial ceiling *may* have been placed on the evaluation of the veracity of magnitude estimates of risk made by the *lay* samples, *if* members of the lay groupings were more likely to view the task of making a "risk" estimate as one of qualitative evaluation.

Since Slovic et al.'s (1985) study of expert-lay differences in risk judgement, several other papers have taken a similar theme. These have used expert samples of toxicologists (Kraus et al., 1992; Slovic et al., 1995), computer scientists (Gutteling and Kuttschreuter, 1999), nuclear scientists (Flynn et al., 1993), aquatic scientists (McDaniels et al., 1997), loss prevention managers in oil and gas production (Wright et al., 2000) and scientists in general (Barke and Jenkins-Smith, 1993). These studies concluded that there are substantial differences in the way that experts and samples of the lay population judge risk. Generally, experts perceive the risks as less than the lay public with regard to the questions asked and the substantive domains. The two exceptions are the studies by Wright et al. (2000) – where experts and members of the lay public shared similarities in risk perception of hazardous events in oil and gas production in the North Sea, and Mumpower et al. (1987), where the rating of the political riskiness of countries by undergraduate students closely paralleled the ratings of professional analysts. Both these sets of results contrast sharply with results of Slovic et al. (1985), described earlier, where the experts saw 26 out of 30 activities/technologies as *more* risky than each of the three lay groupings. However, in all studies, except for the latter study, the relative validity of expert versus lay risk assessments (in terms of the veracity of frequency estimates) *has not* been measured – hence, the commonly accepted view about expert-lay differences in risk judgements rests on the results of a single study that used just 15 experts and which compared their judgements of "risk" with those of groups of lay persons on a task where the validity standard (mortality rates) was not made salient to the lay group. Further, it would seem highly unlikely that the experts who took part in the Slovic et al. study could have had substantive expert knowledge in all of the variety of hazards that were utilized (including mountain climbing, nuclear power and spray cans), which begs the question: Were they truly experts? This might also, in part, explain why the results from this expert sample were inconsistent with the results from expert samples in the other studies. In a review of these studies, Rowe and Wright (2001) concluded that, contrary to received wisdom, there is little empirical evidence for the proposition that experts are more veridical in their risk assessments than members of the public.

More widely, Bolger and Wright (1994) and Rowe and Wright (2001) have argued that in many real-world tasks, apparent expertise (as indicated by, for example, status) may have little relationship to any real judgement skill at the task in question. In Bolger and Wright's review of studies of expert judgemental performance they found that only six had showed "good" performance by experts, while nine had shown poor performance. As we have seen in section 1 of this chapter, Bolger and Wright analysed and then interpreted this pattern of performance in terms of the "ecological validity" and "learnability" of the tasks that were posed to the experts. To reprise, by "ecological validity" we mean the degree to which the experts were required to

make judgements inside the domain of their professional experience and/or express their judgements in familiar metrics. By "learnability" we mean the degree to which it is possible for good judgement to be learned in the task domain. That is, if objective data and models and/or reliable and usable feedback are unavailable, then it may not be possible for a judge in that domain to improve his or her performance significantly with experience. In such cases, Bolger and Wright argued, the performance of novices and "experts" is likely to be equivalent, and they concluded that expert performance will be largely a function of the interaction between the dimensions of ecological validity and learnability – if both are high then good performance will be manifest, but if one or both are low then performance will be poor.

From the perspective of Bolger and Wright's analysis, it is by no means certain that expert risk assessors will be better at judging the veridical risks of hazards than lay persons, and the limited empirical evidence cannot be considered compelling (Rowe and Wright, 2001). This has important implications for the communication of *judgements* of risk. As Rowe and Wright (2001) have argued, in hazard evaluations where the hazardous events happen rarely, if at all, then learnability will be low, and the veridicality of judgements of the magnitude of risks by experts will be suspect. For example, consider the validity of expert predictive judgements about the likelihood magnitude of human infection by "mad cow disease" resulting from eating beef from herds infected with Bovine Spongiform Encephalopathy (BSE) in the early 1990s and the subsequent, poorly predicted, mortality rates (Maxwell, 1999). In this instance, UK politicians selectively used expert predictions to reassure a frightened general public.

Wright et al. (2002) considered the issue of expert-lay differences in frequency, and relative frequency, judgements of lethal events using a sample of professional risk assessors. They extended and developed the study of Lichtenstein et al. (1978) and followed up the suggestion in Shanteau's (1978) commentary on that paper. They utilized a sample of life underwriters, of varying degrees of experience, and a task requiring assessment of a varied set of potentially lethal events.

The results from their study revealed that although both lay and expert groups showed relatively good performance in terms of the ordering of the absolute likelihood (marginal probabilities) and lethality (conditional probabilities) of events, as demonstrated by significant obtained correlations, they also showed similar and systematic bias in terms of overestimating these values. Such overestimation was almost uniform over the hazards for the direct marginal judgements, although less so for conditionals. The student group was no worse at direct marginal or direct conditional estimation than the experts.

Because the *direct* estimation of risks associated with potentially lethal events is an unusual task, even for the experts (at least for marginal estimates, although for conditional estimates the Chief Underwriter stated

that this assessment mode captured the essence of his work-a-day task), we also obtained marginal and conditional estimates in a second, *indirect* way, namely, through pairwise comparisons. Correlational analysis revealed a trend that the experts were indeed better at the task, in terms of identifying which events of the *pairs* led to more deaths (marginals) and were more lethal (conditionals), although these correlations were not significantly different from those of the lay group. However, further analysis revealed that the experts *did* make significantly better judgements than lay person on marginal estimates in terms of ratios (i.e. the number of times one event was more likely to cause death than another) and conditionals (i.e. the number of times an event was more likely to cause death than the other, given that the event happens to someone). In spite of this, both lay persons and experts made the same general errors in the pairwise comparison tasks – namely, in underestimating the ratio of more-to-less ubiquitous and fatal hazards by overly compressing their ranges of estimates.

Section conclusion

As we have reviewed, the evidence for experts being better at the judgements of risks is not strong (see Rowe and Wright, 2001 for a review) and yet has been so readily accepted that there has been no apparent effort to research the topic further. For "true" expertise to be manifest (expertise related to performance, as opposed to social and political imperatives), Bolger and Wright (1994) have argued that the expert must perform a task that is ecologically valid, and the task must also be learnable. Wright et al. (2002) attempted to ensure that their expert-task match was as strong as possible (given experimental limitations), and that ecological validity was high, and yet still obtained expert performance that was not much better than lay person performance. This result suggests that the underwriting task is not truly "learnable", i.e. it is not one for which there is regular feedback on the correctness or otherwise of judgements. Indeed, in the training of underwriters, performance is assessed according to the similarity of junior underwriters' judgements to those of their seniors (Bolger et al., 1989). Once "trained", underwriters receive infrequent performance-related, objective feedback about the correctness of their judgements, and indeed it would be difficult to provide such feedback, given that a "poor" judgement might turn out to be insuring an applicant who subsequently died of a condition after perhaps 20 years of a 25-year policy.

We infer that the tasks performed by *other* professional risk assessors may also be unlearnable. For example, in the case of major hazards in the nuclear industry there may be no risk/judgement feedback at all. From this, we suggest that expert-lay differences in the accuracy of such risk judgements, or in the nature of such judgements (given that the biases evidenced in the Wright et al. (2002) study were similar across lay and expert groups), cannot be assumed. Further, even if experts are significantly more accurate than

lay people, it may still be that differences in accuracy are small, as demonstrated in the present study. Perhaps the commonsense assumption of the superiority of expert risk assessors in making risk judgements is ill-founded. Certainly, future research needs to pay more attention to the *de facto* nature of the learnability of tasks performed by professional risk assessors.

Advice giving and changes in judgement: directions for future research

Our general conclusion is that experts can make valid judgements of probability – if the task conditions are amenable. To make valid judgements, we contend that a prediction/outcome feedback loop must be in place to enable learnability. Also, the judgement task itself should be matched with the expert's knowledge base, and the metric used to elicit probability judgements should be both familiar and acceptable to the expert. Such conditions prevail in meteorologists' predictions of weather events – where excellent calibration is the rule. Similarly, in the domain of risk judgement, such task conditions must, we contend, also prevail for valid judgements to be elicited from experts. So, we conclude that the ideal situation for the expression of valid judgement by experts is when (i) the ecological validity of both the judgement task and the elicitation metric are both high and (ii) when the task itself is truly learnable. In many situations, of course, the second condition may not hold and expert judgement will, we contend, likely be as poor as that of the lay population and subject to similar heuristics and biases. In these situations, there will exist no track-record of prior judgements or associated hit-rate. Here, the judgements made are likely to be one-off or unique. In such situations, the expert has only access to his/her heuristics and the advice of other people – perhaps also experts. The advisors may disagree, so how are/should such disagreements resolved?

Rowe and Wright (1999) studied change in expert opinion amongst members of such expert groups and found that the experts who held the more accurate opinions changed their opinions less than experts who held less accurate opinions over rounds in a mediated group process called "Delphi". One avenue for future research is, therefore, the exchange of knowledge and opinion between experts, but this research area is, as yet, under-explored.

For example, Brockriede and Ehninger (1960) have shown that only a limited number of argument types are, in principle, available to people advocating specific propositions or claims – arguments of parallel case, analogy, motivation and authority.

In *Analogous reasoning*, the reason given makes use of our general knowledge of relationships between two events in dissimilar situations. For example, if someone is trying to estimate the time it will take to drive to a nearby airport, an advisor may reason that "the airport is roughly the same distance

away as the shopping mall. Therefore, the time it will take to get to the airport will be approximately the same as it is to travel to the shopping mall – about 30 minutes."

Parallel case reasoning involves making use of our knowledge of a previous experience of a near identical situation. For example, if someone is trying to estimate the time it will take to drive to a nearby airport an advisor may reason that, "it will take about 30 minutes to drive to the airport because it took me 30 minutes at the same time of day last month."

Authoritative reasoning involves making use of substantive knowledge. For example, "the radio announcer has said that traffic to the airport is heavy today and so I estimate that you should add 20 minutes to your journey time."

Motivational reasoning involves making use of specific insights about people's motivations or desires. For example, "since you will be in a hurry, then I reckon that you can cut five minutes off your usual journey time."

Research has shown that these argument types can be persuasive in some circumstances (McCroskey, 1969; Smith, 1972; Stanchi, 2006), but outside a legal context, no research has been conducted to explore the persuasiveness effects of different forms of argument structure on opinion change in experts. Thus, many crucial questions remain to be explored and answered. For example, what components of advice-giving cause opinion-change in individual experts? How is advice evaluated and under what conditions will advice be assimilated or discounted? When one expert defers in his or her own opinion to the well-argued opinion of another, is this an indicator of the presence of valid advice that will improve validity in (the revised) judgemental prediction? In our view, the study of the use of argument in advice-giving will become a major topic in investigations focused on understanding and improving expert judgement of probability and risk in unique, or one-off, situations where the expert cannot utilize learnability, and the expert is aware that his/her own judgement may be influenced by inappropriate heuristics that could lead to bias.

References

Ayton P and McClelland AGR. (1997). How real is overconfidence? *Journal of Behavioral Decision Making* **10**:279–285.

Barke RP and Jenkins-Smith HC. (1993). Politics and scientific expertise: scientists, risk perception, and nuclear waste policy. *Risk Analysis* **13** (4): 425–439.

Bolger F and Wright G. (1994). Assessing the quality of expert judgement: issues and analysis. *Decision Making Systems* **11**:1–24.

Bolger F, Wright G, Rowe Gammack J, and Wood RJ. (1989). Lust for life: developing expert systems for life assurance underwriting. In: Shadbold N (ed.) *Research and Development in Expert Systems, VI*. Cambridge University Press: Cambridge.

Brockriede W and Ehninger D. (1960). Toulmin on argument: an interpretation and application. *Quarterly Journal of Speech* **46**:44–53.

Carmines EG and Zeller RA. (1979). *Reliability and Validity Assessment*. Sage University Papers Series: Beverley Hills, CA.
Compte O and Postlewaite A. (2004). Confidence-enhanced performance. *American Economic Review* **94**:1536–1557.
Dalkey N. (1972). An elementary cross-impact model. *Technological Forecasting and Social Change* **3**:341–351.
Dougherty MRP, Getys CF, and Ogden EE. (1999). MINERVA-DM: a memory processes model for judgements of likelihood. *Psychological Review* **106**:180–209.
Dube-Rioux L and Russo JE. (1988). An availability bias in professional judgement. *Journal of Behavioural Decision-Making* **1**:233–237.
Ducharme WM and Peterson CR. (1968). Intuitive inference about normally distributed populations. *Journal of Experimental Psychology* **78**:269–275.
Eddy DM. (1982). Probabilistic reasoning in clinical medicine: problems and opportunities. In: Kahneman D, Slovic P, and Tversky A (eds). *Judgement Under Uncertainty: Heuristics and Biases*. Cambridge University Press: Cambridge.
Edwards W, Phillips LD, Hays WL, and Goodman BC. (1968). Probabilistic information processing systems. *IEEE Transactions on Systems Science and Cybernetics* **4**:248–265.
Einhorn HJ (1980) Learning from experience and sub-optimal rules in decision making. In Wallsten T (ed.) Cognitive Processes in Choice and Decision Behavior. Hillsdale, N.J.: Erlbaum.
Einhorn HJ and Hogarth RM. (1978). Overconfidence in judgement. Persistence of the illusion of validity. *Psychological Review* **85**:394–476.
Ferrell WR and McGoey PJ. (1980). A model for calibration of subjective probabilities. *Organizational Behavior and Human Decision Processes* **26**:32–53.
Flynn J, Slovic P, and Mertz CK. (1993). Decidedly different: expert and public views of risks from a radioactive waste repository. *Risk Analysis* **13** (6): 643–648.
Goodwin P and Wright G. (2004). *Decision Analysis for Management Judgement*. Wiley: Chichester.
Griffin D and Tversky A. (1992). The weighting of evidence and the determinants of overconfidence. *Cognitive Psychology* **24**:411–435.
Gutteling JM and Kuttschreuter M. (1999). The millennium bug controversy in the Netherlands? Experts' views versus public perception. In: Goossens LHJ (ed.) *Proceedings of the 9*th *Annual Conference of Risk Analysis: Facing the Millennium*. Delft University Press: Delft, Netherlands, pp. 489–493.
Hoerl A and Fallin HK. (1974). Reliability of subjective evaluation in a high incentive situation. *Journal of the Royal Statistical Society* **137**:227–230.
Kabus I. (1976). You can bank on uncertainty. *Harvard Business Review*, May–June, **54**:95–105.
Kahneman D, Slovic P, and Tversky A. (1982). *Judgement Under Uncertainty: Heuristics and Biases*. Cambridge University Press: Cambridge.
Kanal LN and Lemmer JF (eds). (1986). *University and Artificial Intelligence*. Elsevier: Amsterdam.
Keren G. (1987). Facing uncertainty in the game of bridge: a calibration study. *Organisational Behaviour and Human Decision Processes* **39**:98–114.
Keren G. (1990). Cognitive aids and debiasing methods. In: Caverni JP et al. (eds). *Cognitive Biases*. Elsevier: Amsterdam.
Kraus N, Malmfors T, and Slovic P. (1992). Intuitive toxicology: expert and lay judgements of chemical risks. *Risk Analysis* **12** (2): 215–232.
Lemmer JF and Kanal LN (eds). (1988). *Uncertainty and Artificial Intelligence 2*. Elsevier: Amsterdam.

Lichtenstein S, Fischhoff B, and Phillips LD. (1982). Calibration of probabilities: the state of the art to 1980. In: Kahneman D, Slovic P, and Tversky A (eds). *Judgement Under Uncertainty: Heuristics and Biases*. Cambridge University Press: New York.

Lichtenstein S, Slovic P, Fischoff B, Layman M, and Combs B. (1978). Judged frequency of lethal events. *Journal of Experimental Psychology: Human Learning and Memory* **4**:551–578.

Maxwell RJ. (1999). The British government's handling of risk: some reflections on the BSE/CJD crisis. In: Bennett P and Calman K (eds). *Communications and Public Health*. Oxford University Press: Oxford, pp. 94–107.

McClelland AGR and Bolger F. (1994). The calibration of subjective probabilities: theories and models 1940–94. In: Wright G and Ayton P (eds). *Subjective Probability*. Wiley: Chichester.

McCroskey JC. (1969). Toward an understanding of the importance of "evidence" in persuasive communication. *The Pennsylvania Speech Annual* **23**:65–71.

McDaniels TL, Axelrod LJ, Cavanagh NS, and Slovic P. (1997). Perception of ecological risk to water environments. *Risk Analysis* **17** (3): 341–352.

Mumpower JL, Livingston S, and Lee TJ. (1987). Expert judgments of political riskiness. *Journal of Forecasting* **6**:51–65.

Murphy AH and Brown BG. (1985). A comparative evaluation of objective and subjective weather forecasts in the United States. In: Wright G (ed.) *Behavioural Decision Making*. Plenum: New York.

Oskamp S. (1965). Overconfidence in case-study judgements. *Journal of Consulting Psychology* **29**:261–265.

Phillips LD and Edwards W. (1966). Conservatism in a simple probabilistic inference task. *Journal of Experimental Psychology* **72**:346–354.

Phillips LD, Hays WL, and Edwards W. (1966). Conservatism in complex-probabilistic inferences. *IEEE Transactions on Human Factors in Electronics* **7**:7–18.

Rowe G and Wright G. (1999). The Delphi technique as a forecasting tool: issues and analysis. *International Journal of Forecasting* **15**:353–375.

Rowe G and Wright G. (2001). Differences in experts and lay judgements of risk: myth or reality? *Risk Analysis* **21**:341–356.

Schafer RE, Borcherding K, and Laemmerhold CL. (1977). Consistency of future event assessments. In: Jungermann H and de Zeeuw G (eds). *Decision Making and Change in Human Affairs*. Reidel, Dordrecht: The Netherlands.

Shafer G. (1987). Probability judgement in artificial intelligence and expert systems. *Statistical Science* **2**:3–44.

Shanteau J. (1978). When does a response error become a judgement bias? Commentary on "judged frequency of lethal events". *Journal of Experimental Psychology: Human Learning and Memory* **4** (6): 579–581.

Shanteau J. (1992). The psychology of experts: an alternative view. In: Wright G and Bolger F (eds). *Expertise and Decision Support*. Plenum: New York.

Slovic P. (1987). Perception of risk. *Science* **236**:280–285.

Slovic P. (1999). Trust, emotion, sex, politics and science: surveying the risk-assessment battlefield. *Risk Analysis* **19** (4): 689–701.

Slovic P, Fischhoff B, and Lichtenstein S. (1985). Characterizing perceived risk. In: Kates RW, Hohenemser C, and Kasperson JX (eds). *Perilous Progress: Managing the Hazards of Technology*. Westview: Boulder, CO, pp. 91–125.

Slovic P, Malmfors T, Krewski D, Mertz CK, Neil N, and Bartlett S. (1995). Intuitive toxicology II. Expert and lay judgements of chemical risks in Canada. *Risk Analysis* **15** (6): 661–675.

Smith TJ. (1972). *The Effects of Truth and Desirability Evidence on Judgements of Truth and Desirability of a Proposition*. Unpublished Master's thesis. Michigan State University: East Lansing.

Stanchi KM. (2006). The science of persuasion: an initial exploration. *Michigan State Law Review I* **52**:1–45.

Tonn BE, Goeltz RT, and Travis C. (1992). Eliciting reliable uncertainty estimates. *Expert Systems* **9** (1): 25–33.

Tversky A and Koehler DJ. (1994). Support theory: a nonextensional representation of subjective-probability. *Psychological Review* **101**:547–567.

Wagenaar WA and Keren GB. (1986). Does the expert know? The reliability of predictions and confidence ratings of experts. In: Hollnagel E et al. (eds). *Intelligent Decision Support in Process Environment*. Springer-Verlag: Berlin.

Wallace HA. (1923). What is the corn Judge's mind? *Journal of the American Society of Agronomy* **15**:300–304.

Wallsten TS and Budescu DV. (1983). Encoding subjective probabilities: a psychological and psychometric review. *Management Science* **29**:151–173.

Wright G and Ayton P. (1987). *Judgemental Forecasting*. Wiley: Chichester.

Wright G, Bolger F, and Rowe G. (2002). An empirical test of the relative validity of expert and lay judgments of risk. *Risk Analysis* **22** (6): 1107–1122.

Wright G, Pearman A, and Yardley K. (2000). Risk perception in the UK oil and gas production industry: are expert loss-prevention managers' perceptions different from those of members of the public? *Risk Analysis* **20**:681–690.

Wright G, Rowe G, Bolger F, and Gammack G. (1994). Coherence, calibration and expertise in judgemental probability forecasting. *Organisational Behavior and Human Decision Processes* **57**:1–25.

Wright G, Saunders C, and Ayton P. (1988). The consistency, coherence and calibration of holistic, decomposed and recomposed judgemental probability forecasts. *Journal of Forecasting* **7**:185–199.

Yates JF. (1990). *Judgement and Decision Making*. Prentice-Hall: Englewood Cliffs, NJ.

Youseff ZI and Peterson CR. (1973). Intuitive cascaded inferences. *Organisational Behaviour and Human Performance* **10**:349–358.