Risk Analysis, Vol. 22, No. 6, 2002

An Empirical Test of the Relative Validity of Expert and Lay Judgments of Risk

George Wright,[1]* Fergus Bolger,[2] and Gene Rowe[3]

This article investigates how accurately experts (underwriters) and lay persons (university students) judge the risks posed by life-threatening events. Only one prior study (Slovic, Fischhoff, & Lichtenstein, 1985) has previously investigated the *veracity* of expert versus lay judgments of the magnitude of risk. In that study, a heterogeneous grouping of 15 experts was found to judge, using marginal estimations, a variety of risks as closer to the true annual frequencies of death than convenience samples of the lay population. In this study, we use a larger, homogenous sample of experts performing an ecologically valid task. We also ask our respondents to assess frequencies and relative frequencies directly, rather than ask for a "risk" estimate—a response mode subject to possible qualitative attributions—as was done in the Slovic *et al.* study. Although we find that the experts outperformed lay persons on a number of measures, the differences are small, and both groups showed similar global biases in terms of: (1) overestimating the likelihood of dying from a condition (marginal probability) and of dying from a condition given that it happens to you (conditional probability), and (2) underestimating the ratios of marginal and conditional likelihoods between pairs of potentially lethal events. In spite of these scaling problems, both groups showed quite good performance in ordering the lethal events in terms of marginal and conditional likelihoods. We discuss the nature of expertise using a framework developed by Bolger and Wright (1994), and consider whether the commonsense assumption of the superiority of expert risk assessors in making magnitude judgments of risk is, in fact, sensible.

1. INTRODUCTION

In a pioneering paper, Lichtenstein, Slovic, Fischhoff, Layman, and Combs (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. These investigators argued that "Citizens must assess risks accurately in order to mobilize society's resources effectively for reducing hazards .... official recognition of the importance of valid risk assessments is found in the 'vital statistics' that are carefully tabulated ... there is, however, no guarantee that these statistics are reflected in the public's intuitive judgments" (1978:551). 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 individual events'

[1] Graduate School of Business, University of Strathclyde, 199 Cathedral Street, Glasgow G4 0QU, UK.
[2] Department of Management, Bilkent University, 06533 Bilkent, Ankara, Turkey.
[3] Institute of Food Research, Norwich Research Park, Colney, Norwich NR4 7UA, UK.
*Address correspondence to George Wright, Graduate School of Business, University of Strathclyde, 199 Cathedral Street, Glasgow G4 0QU, UK.

0272-4332/02/1200-1107$22.00/1

frequencies were poor.[4] 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 judgment of lethal events using subjects who have direct knowledge and exposure to such events (such as life insurance analysts)" (1978:581).

Since the 1978 paper, research on risk judgments has led to the generally accepted conclusion that expert judgments 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, Fischhoff, and Lichtenstein (1985). In this study, the authors utilized samples of the U.S. League of Women Voters, university students, members of the U.S. Active Club (an organization of business and professional people devoted to community services activities), and a group of "experts" (comprising 15 people who were described as professional risk assessors, including a geographer, an environmental policy analyst, an economist, a lawyer, a biologist, and a government regulator of hazardous materials). 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" (1985:116)). 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 frequency or relative frequency estimate (cf. Lichtenstein *et al.*, 1978), the avenue was open—for both experts and nonexperts—for qualitative risk attributes, such as the voluntariness or controllability of the risk, to enter into these global risk judgments.

Slovic, Fischhoff, and Lichtenstein (1985) concluded that the judgment of their experts differed substantially from nonexpert judgment 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, Fischhoff, and Lichtenstein (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" (1985:95). 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 three lay samples. However, Slovic, Fischhoff, and Lichtenstein (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 & 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, Fischhoff, and Lichtenstein's (1985) study of expert-lay differences in risk judgment, several other papers have taken a similar theme. These have used expert samples of toxicologists (Kraus, Malmfors, & Slovic, 1992; Slovic, *et al.*, 1995), computer scientists (Gutteling & Kuttschreuter, 1999), nuclear scientists (Flynn, Slovic, & Mertz, 1993), aquatic scientists (McDaniels *et al.*, 1997), loss prevention managers in oil and gas production (Wright, Pearman, & Yardley, 2000), and scientists in general (Barke & 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 both across the wide variety of questions asked and across the variety of substantive domains. The two exceptions are the studies by Wright, Pearman, and Yardley (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, Livingston, and Lee (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

4 However, these authors did not provide a standard of evaluation that allowed the implications of this finding for subsequent decision making to be determined. For example, would an individual's personal decisions about exposure to particular hazardous events be sensitive to low accuracy in that individual's assessment of their frequency?

Slovic, Fischhoff, and Lichtenstein (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 judgments rests on the results of a single study that used just 15 experts and that compared their judgments 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 (e.g., a geographer, lawyer, and economist) 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 expert? 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 judgmental skill at the task in question. In Bolger and Wright's review of studies of expert judgmental performance, they found that only six had showed "good" performance by experts, while nine had shown poor performance and the remaining five showed equivocal performance. Bolger and Wright analyzed and then interpreted this pattern of performance in terms of the "ecological validity" and "learnability" of the tasks that were posed to the experts. By "ecological validity" is meant the degree to which the experts were required to make judgments inside the domain of their professional experience and/or express their judgments in familiar metrics. By "learnability" is meant the degree to which it is possible for good judgment to be learned in the task domain. That is, if objective data and models and/or reliable and usable feedback is unavailable, then it may not be possible for a judge in that domain to improve his or her performance significantly with experience (see, e.g., Einhorn, 1980; Keren, 1987). In such cases, Bolger and Wright argued, the performance of novices and "experts" is likely to be equivalent. Bolger and Wright 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. For example, as pointed out by Murphy and Brown (1985), weather forecasters in the United States and other countries are routinely required to give confidence estimates attached to their forecasts. Such forecasts, say for the next day's weather, are succeeded by timely, unconfounded feedback since very few interventions by man can confound the prediction/outcome relationship. It follows that experimental studies of weather forecasters' confidence in their weather predictions will tend to be ecologically valid and that experimental tasks will, most likely, use potential weather events where the participating forecasters have experienced prior conditions of learnability. Indeed, weather forecasters do show good judgment (see Murphy and Brown, 1985, for a review).

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 & Wright, 2001). This has important implications for the communication of *judgments* 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 judgments of the magnitude of risks by experts will be suspect. For example, consider the validity of expert predictive judgments about the likelihood magnitude of human infection by "mad cow disease" resulting from eating beef from herds infected with Bovine Spongiform Encephalopathy in the early 1990s and the subsequent, poorly predicted, mortality rates (Maxwell, 1999). In this instance, UK politicians used expert predictions to inappropriately reassure a frightened general public.

The present study considers the issue of expert-lay differences in frequency, and relative frequency, judgments of lethal events using a sample of professional risk assessors. It extends and develops the study of Lichtenstein *et al.* (1978) and follows up the suggestion in Shanteau's (1978) commentary on that paper. We utilize a sample of life underwriters, of varying degrees of experience, and a task requiring assessment

of a varied set of potentially lethal events. In the next two sections, we attempt to demonstrate that the expert-task relationship in this study has "ecological validity," and hence, that inadequacies in expert performance (if found) may reasonably be attributed to (lack of) expert performance and not to poor task design. First, we describe the nature of the experts used in this study and the tasks they perform professionally and, second, we detail the nature of the judgments that we elicited from our subjects.

2. THE NATURE OF OUR EXPERTS

The details of the underwriting process that took place at the insurance company from which underwriters were recruited in the present study are described in Bolger, Wright, Rowe, Gammack, and Wood (1989). Essentially, financial advisors would send applications for life insurance to local branches of the company. If the applications were straightforward (i.e., applicants were of an appropriate age, the sum to be insured was within reasonable limits, and no unusual circumstances were described on the form), then the branch would forward these to the new business department (NBD) to issue the policy. The decisions by the clerical staff at the branches were made on the basis of a manual provided by the head office that outlined the factors that made a proposal acceptable or unacceptable. If the application was not straightforward, the application would be forwarded to the underwriters at the NBD for them to make the decision. The underwriters might then issue a policy, or reject the application or, often, send for additional information, either medical information from a doctor, or personal information from the applicant (which might, for example, relate to a person's hobbies or intended travel itineraries). See Rowe and Wright (1993) for a discussion of the life-underwriting task and the extent to which underwriting judgment can be modeled in automated expert systems.

At the time of Bolger *et al.*'s paper (1989), approximately 60% of proposals were accepted at branch level, and 40% were sent to the NBD underwriters for assessment. The underwriters sought medical evidence on approximately 20% of the cases they received. The majority of the underwriters' work is based on the internalization of routines. Specifically, life underwriters assess the information they receive from a standard application form, in combination with any additional information that may have been requested (such as medical reports from the applicants' doctors or medical examinations taken for the purpose of the application). The task of the underwriter is to match the applicant to the particular mortality table that correctly predicts the statistical probability of the individual succumbing to death over the term of the policy.

On the basis of the comparison of the applicant to the statistical norm, the underwriter decides whether the application is accepted, declined, or accepted but with additional premiums or waivers of coverage for certain conditions. The skill of underwriters appears to be in the internalization of key heuristics about risk that enables them to rapidly scan application forms for important phrases or indicators, only then referring to the manuals or mortality tables that they have available. More senior underwriters (who are able to underwrite larger values of sum assured) tend to assess applications much more rapidly than their juniors because they do not need to refer to the manuals or tables very often.

3. THE NATURE OF THE TASK

In this study, we ask for marginal and conditional assessments for a variety of hazardous events.[5] Marginal assessments include answers to questions such as "What is the death rate per 100,000 from asthma?" This equates to an assessment of the "risk" associated with a hazard as widely understood by risk assessors. Conditional assessments include answers to questions such as "What is the probability of death from stomach cancer *given that* an individual is diagnosed with the condition?" The exact detail of the presentation of the lethal events to our samples of underwriters and students is given in Section 4.1. In this study, we not only ask for the traditional "risk" figure itself, but also for the conditional component, since we anticipated that life underwriters would be more accurate in their conditional assessments than lay persons because a large proportion of their jobs—especially for more senior underwriters—is that of judgmentally assessing life proposals from those individuals with *existing* medical conditions. That is, application forms detail conditions that applicants *already have*, and underwriters effectively judge the risk of the

[5] These hazardous events were derived from those used by Lichtenstein *et al.* (1978), although various changes had to be incorporated due to: (1) certain of the original U.S.-based items being inappropriate in the United Kingdom (e.g., tornadoes), and (2) the need for base-line data for the calculation of conditional probabilities (e.g., number of hospital admissions, notifications of diseases, reported accidents/crimes, etc.).

applicants dying from those conditions before the term of their policy. Nevertheless, we anticipated that the underwriters would also be more accurate than lay persons for marginal risk assessments (though less so than for the conditional assessments), since the conditional assessment makes up a part of this statistic, and since underwriters should have some familiarity with general mortality tables. (Marginal assessments are usually made by actuaries in compiling "life tables" and as such, it would be interesting in a future study to consider the relative performance of actuaries to underwriters on the different components of risk.)

Following the procedure of Lichtenstein *et al.* (1978), we asked respondents to use two different response modes in assessing the marginal and conditional risks associated with hazards: first, direct assessments, requiring subjects to directly estimate the required statistics; and second, indirect estimates, requiring subjects to make paired comparisons of hazardous events. That is, for pairings of potentially lethal events, subjects were asked: "Which of the two events is the most likely to cause death?" and then "How many times more likely is the event you circled to cause death than the other event in the same question?" We did this for both marginal and conditional assessments—in the latter case, our first question asked "Which of the two events do you think is the most likely to cause death, given it happens to someone?" and the second again asked "How many times more likely (is this)?"

Overall, we reasoned that utilization of direct and indirect methods of risk assessment over a varied set of 31 lethal events should allow us to test if life underwriters were, in fact, more veridical in their risk assessments than a sample of the lay population represented by university students. The nature of the response mode is another important feature with regard to the "ecological validity" of the task (Bolger & Wright, 1994), in that it should accurately match the mode used by the experts in their everyday work.

3.1. Were Our Tasks Ecologically Valid?

Did the tasks that we devised have ecological validity for the underwriters? Does our particular decomposition of risk judgments correspond to the day-to-day work of the experts? To answer these questions we constructed a questionnaire that, having listed the 31 potentially lethal events, provided six exemplar question types from our main study (i.e., a paired marginal/most likely judgment; a paired marginal/times more likely judgment; a paired conditional/most likely judgment; a paired conditional/times more likely judgment; a direct marginal judgment; a direct conditional judgment) and then asked if the respondents would make more accurate assessments than a typical university undergraduate (indicated by a "1" on a seven-point scale) or be no more accurate than a typical university undergraduate (indicated by a "7"). For all question types, the Chief Underwriter of the life insurance company indicated "1." We also asked which of each of the question types characterized her day-to-day work as an underwriter on a scale from "completely" (indicated by a "1") to "not at all" (indicated by a "7"). For all question types the Chief Underwriter indicated "6," with the exception of the direct conditional assessments, where she indicated "1." Finally, we asked the Chief Underwriter, "In assessing an application for life insurance, is your decision reached on the basis of using judgment alone?" This assessment was made using a seven-point scale from "completely" (indicated by a "1") to "not at all" (indicated by a "7"). The box ticked for this assessment was "2."

We did ask the Chief Underwriter to distribute the questionnaire to her subordinates but she could not see the point of doing so. In an effort to validate this response we interviewed the Development Underwriter of a competitor life insurer. A Development Underwriter's job is to establish the underwriting guidelines for underwriting manuals. This respondent, too, indicated that underwriters should be able to make more accurate assessments than typical university undergraduates (indicated by a "1" on each of our questions on this issue). When asked the question regarding the relative use of judgment in risk assessment, this respondent indicated "4." This respondent, too, could not see the point of distributing the questionnaire to his company's underwriters since it was "obvious that underwriters would perform better on the risk assessments than undergraduates."

Since our study asked respondents to give likelihood and relative likelihood judgments, we expected that both our experts and our students would address this task directly, rather than as in the Slovic, Fischhoff, and Lichtenstein (1985) study, described earlier, where the nonexperts, perhaps, focused more than the experts on *qualitative risk attributes* in generating the overall risk judgments required by the experimental procedure. As such, our response mode is a more direct assessment of nonexperts' ability to produce veridical judgments of actuarial risk—where experts had shown very good judgments in the Slovic, Fischhoff, and Lichtenstein (1985) study.

Table I. Median Direct Marginal Assessments of Students and Underwriters Compared to True Marginal Values

Lethal Events	Students	Underwriters	True Marginals
Lung cancer	175.00	300.00	71.200
Breast cancer	150.00	150.00	53.100
All accidents	175.00	90.00	25.340
Bronchitis/emphysema	150.00	200.00	24.200
Stomach cancer	75.00	117.00	19.980
Diabetes	127.50	200.00	15.750
Smoking	238.00	250.00	13.790
Domestic accident	67.50	15.00	11.570
Car accident	50.00	125.00	10.000
Hypertension	40.00	22.00	8.350
Leukemia	15.00	6.36	7.130
Road accident	150.00	25.00	5.680
Alcohol/drugs	125.00	310.00	4.840
Asthma	5.50	30.00	3.970
Benign tumor	27.00	0.40	2.730
Suffocation/choking	45.00	30.00	1.720
Fire and flames	20.00	30.00	1.290
Accidental poisoning	75.00	50.00	1.280
Hodgkin's disease	10.00	80.00	0.970
Meningitis	3.50	0.89	0.490
Food poisoning	100.00	5.00	0.390
Rail accident	11.00	5.00	0.240
Infectious hepatitis	115.00	100.00	0.220
Electric shock	40.50	0.35	0.220
Pregnancy/childbirth	5.00	10.00	0.110
Salmonella	3.00	0.90	0.110
Air accidents	11.00	5.00	0.060
Measles	1.50	5.00	0.020
Lightning	1.00	0.003	0.010
Whooping cough	60.00	100.00	0.010
Polio	1.75	5.00	0.002

Table II. Direct Conditional Assessments

Lethal Events	Students	Underwriters	True Conditional Probability
Stomach cancer	0.50000	0.50000	0.83000
Air accidents	0.95000	0.94500	0.72000
Lung cancer	0.80000	0.39000	0.70000
Smoking	0.25500	0.40000	0.56000
Bronchitis/emphysema	0.20000	0.32500	0.55000
Polio	0.17500	0.00400	0.33000
Hypertension	0.12500	0.00500	0.31000
Breast cancer	0.30000	0.57500	0.30000
Leukemia	0.50000	0.29300	0.25000
Lightning	0.50000	0.89000	0.20000
Diabetes	0.10000	0.10000	0.09500
Meningitis	0.60000	0.00300	0.08900
Hodgkin's disease	0.55000	0.49500	0.06900
Infectious hepatitis	0.40000	0.00750	0.06800
Fire and flames	0.20000	0.20000	0.06000
Rail accident	0.30000	0.25000	0.05300
Domestic accident	0.11000	0.00150	0.04600
Salmonella	0.00800	0.00040	0.03200
Road accident	0.35000	0.00500	0.02900
All accidents	0.10000	0.10250	0.02800
Asthma	0.00300	0.10000	0.02800
Alcohol/drugs	0.25000	0.25000	0.02000
Food poisoning	0.00600	0.00275	0.02000
Suffocation/choking	0.47500	0.27500	0.01800
Car accident	0.14500	0.15000	0.01700
Electric shock	0.00650	0.00250	0.01600
Benign tumor	0.00700	0.00200	0.01500
Accidental poisoning	0.27000	0.00600	0.00600
Pregnancy/childbirth	0.00100	0.00065	0.00040
Measles	0.00075	0.00050	0.00010
Whooping cough	0.40500	0.00250	0.00008

4. METHOD

4.1. Procedure

Our questionnaire utilized 31 hazardous events[6] divided into three blocks of eight and one block of seven. Tables I and II list the events and give the "true" marginal rates and conditional probabilities.[7] In our procedure, we constructed four blocks of events, where each event was (roughly) identified as high or low on the scales of true marginals and conditionals. The events allocated to each block were allocated in order to match the permutations of such scalings.

Each respondent to our questionnaire responded to only one block of lethal events and made four forms of assessments: direct marginal, direct conditional, paired marginal, and paired conditional. We wanted complete pairwise comparisons of events (in order to test for consistency and uncover the subjective scale for all events). However, this procedure results in a combinatorial explosion of comparisons. In order that the number of comparisons to be made was not too onerous for the participants, we decided to

6 The full list actually comprised 32 events. One event, labeled "homicide," was meant to address "intended homicide," as was indicated by the notes attached to the questionnaire. However, a number of subjects clearly interpreted this item as actual homicide, giving a conditional probability rating of 1.0, which by definition is correct! Subjects in all conditions actually completed the same number of ratings, though we decided to omit responses to the (differentially interpreted) homicide item from all analysis.

7 The "true" values are statistical estimates and are influenced by the vagaries of reporting, classification, attribution, etc. In the instructions to questionnaires (given in Appendices 1, 2, and 3), we attempted to communicate official definitions and practices to our respondents. Obviously, we, along with Lichtenstein *et al.* (1978), cannot be sure of the degree to which our respondents, especially our student grouping, absorbed the details of the definitions that they were given.

use only eight events per participant (or seven events for those participants who responded to the smaller block of events), which resulted in 28 (or 21) marginal and 28 (or 21) conditional pairwise comparisons for each respondent. The presentation of sets of marginal and conditional pairings was counterbalanced and the ordering of pairings was randomized. The two sets of pairings were always followed by the set of direct assessments.

The direct assessments were made on a single side of paper with the events listed on the left-hand side and with two columns. The first column required marginal assessments, the second conditional. Counterbalanced low and high anchor points were given, randomly, as examples to prompt these direct estimates: heart disease (marginal rate = 326/100,000, conditional probability = 0.47) and industrial accident (marginal rate = 1/100,000, conditional probability = 0.003).

Appendix 1 contains the instructions given to the paired marginal questions, Appendix 2 contains the instructions for the paired conditional questions, and Appendix 3 contains the instructions for the direct assessment questions. The instructions for the paired marginal questions and the direct marginal assessments were similar to those used by Lichtenstein *et al.* (1978). Our sets of conditional questions (not studied by Lichtenstein *et al.*) were derivations of the instructions given to the marginal questions.

4.2. Respondents

Thirty-seven life underwriters completed the questionnaire. They were all from one life insurance company. Thirty-nine business students at Bristol Polytechnic also completed the questionnaire. Approximately equal numbers of respondents received each of the four blocks of stimulus items.

5. RESULTS

5.1. Direct Marginal Assessments

Table I sets out the median responses of our underwriters' and students' marginal assessments on the third section of our questionnaire—the direct marginal estimates. The table also gives the true marginal rates per 100,000 for these hazardous events and is ordered by these true rates.

As can be seen, for all 31 events, our students' median assessment was higher than the true marginal rate. For the underwriters, only three of their 31 median assessments were lower than the true marginal

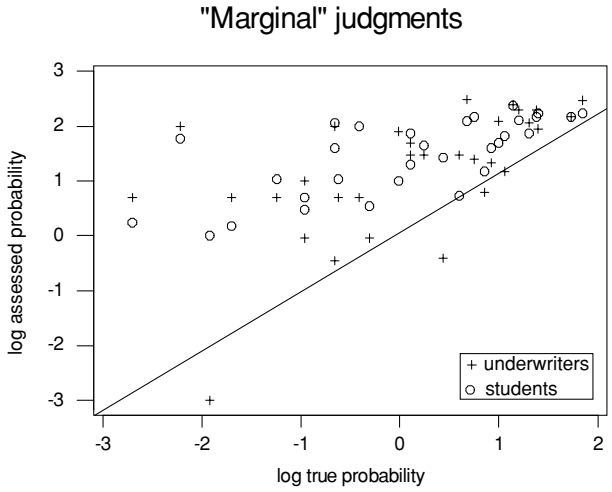

Fig. 1. Log of directly assessed "marginal" death rates for the 31 events plotted against log-true marginal rates for both students and underwriters. Perfect assessments fall on the diagonal identity line.

rates. We correlated the median assessments from the two samples with the true marginal rates to see whether the underwriters' assessments had a stronger relationship with the true rates than did the students'. The rank-ordered correlation between the students' assessment and the true rates was 0.73 ($p < 0.01$), whereas the correlation for our underwriters was 0.66 ($p < 0.01$), indicating that the students were, against expectation, as veridical in their ordering of risk assessments as the underwriters since the difference between the correlations was not significant. Fig. 1 graphs the logarithm of the true marginal rates against the logarithm of the median marginal assessments for both the student and underwriter samples. Generally, both samples overestimate the risk for the potentially life-threatening events.

5.2. Direct Conditional Assessments

Table II sets out the median responses of the underwriters' and students' direct conditional assessments. The table also gives the true conditional probabilities and is ordered by these true probabilities.

Only two of the median assessed conditionals were roughly veridical: the students' assessments of breast cancer and the underwriters' assessments of accidental poisoning. Of the remaining assessments, 65% of the students' and 52% of the underwriters' assessments were greater than the true conditional probability. The rank-ordered correlation between the students' median assessments and the true conditionals was 0.53 ($p < 0.01$), whereas for the

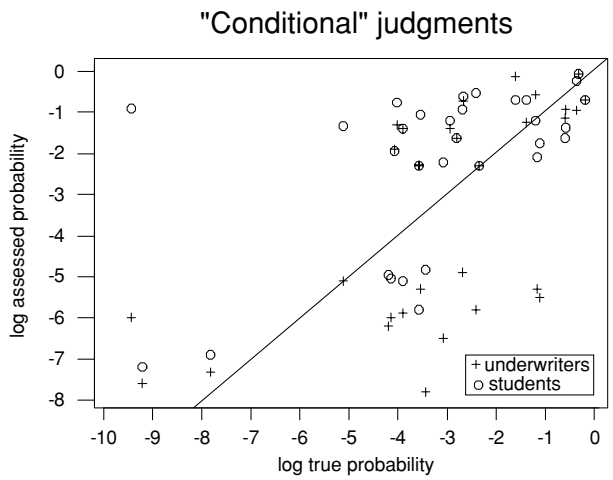

Fig. 2. Log of directly assessed "conditional" death rates for the 31 events plotted against log-true conditional rates for both students and underwriters. Perfect assessments fall on the diagonal identity line.

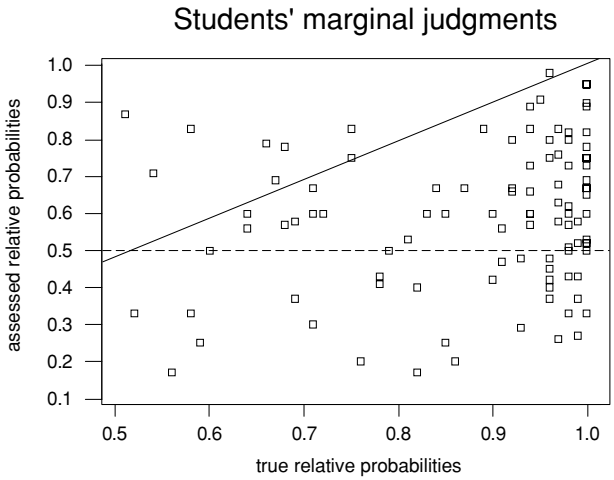

Fig. 3a. Students' 105 pairwise assessments of "marginal" death rates each expressed as a probability that the truly most likely event of each pair will lead to death relative to the truly least likely event of each pair. These assessed relative probabilities are plotted against the true relative probabilities calculated in the same manner. Perfect assessments fall on the diagonal identity line and correct identifications of the truly most likely event of each pair fall above the dashed horizontal line.

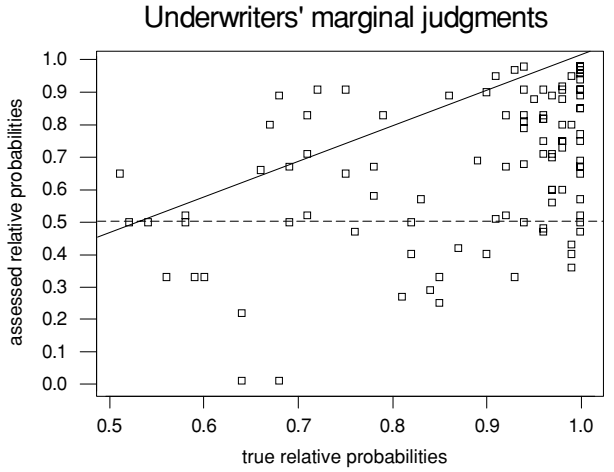

Fig. 3b. Underwriters' 105 pairwise assessments of "marginal" death rates each expressed as a probability that the truly most likely event of each pair will lead to death relative to the truly least likely event of each pair. Assessed relative probabilities are plotted against the true relative probabilities calculated in the same manner. Perfect assessments fall on the diagonal identity line and correct identifications of the truly most likely event of each pair fall above the dashed horizontal line.

underwriters it was 0.66 ($p < 0.01$). Although the underwriters' performance was better than the students' on this aspect, these correlations were not found to be significantly different from one another. Fig. 2 graphs the logarithm of the true conditional probabilities against the logarithm of the median conditional probability assessments. Unlike the marginal assessments, there is no systematic bias of overestimation: the probability of some events was underestimated and others overestimated.

5.3. Paired Assessments: Group Marginal Analysis

Fig. 3a graphs, on the y-axis, the students' median marginal probabilities for each of the pairings of lethal events within each of the four groupings of events, i.e., there are 105 data points. These medians are for the assessed relative likelihood of death from the events, converted to half-range probabilities (i.e., 0.5 to 1.0), of the truly most likely event of a pair occurring. This value will approach 1.0 when "times more likely the event you circled causes death" approaches infinity. For example, if the "times more likely" is 99, then the probability is 0.99. If the "times more likely" is thought to be 300, then the probability is 300 divided by 301, which equals 0.99667, and so on. Values that fall below 0.5 on the y-axis represent cases in which, on average, the respondents incorrectly identify the more likely cause of death. The number of dots above this line, in proportion to the number of dots below, shows the proportion of median responses that correctly identified the truly most likely cause of death from the pairing. On the x-axis, we plot the true half-range probability of each truly more likely event occurring, calculated in the same manner (such that all values lie between 0.5 and 1.0). Fig. 3b repeats this

analysis for the underwriters. This analysis is analogous to that shown in Lichtenstein *et al.*'s (1978). Figs. 2 and 3, with the exception that their "true ratio" x-axis is compressed here due to the half-range conversion of the relative likelihoods. Note that the figures from Lichtenstein *at al.* (1978) plot true ratio against percentage correct.

As can be seen by visual inspection of the two figures, both students and underwriters generally underestimate the relative likelihoods of the events, in that the relative median assessed likelihoods are generally lower than the true relative likelihoods (as indicated by the occurrence of most of the points below the diagonal in both figures). In other words, both groups of respondents do not appreciate the true (larger) differences in rates. This is particularly so in cases where the true probability is high (that is, the ratio in terms of number of deaths between the two lethal events is high). In short, both groups of respondents appear to be using compressed ranges of ratio estimates, resulting in over-homogenous estimations. If we consider the estimates that fall below a horizontal line drawn at 0.5 on the y-axis, then this indicates occasions in which the wrong option of a pair was, on average, estimated to be the more likely cause of death. Application of chi-square tests on a two-by-two contingency table, which was made by counting the number of points above and below the horizontal line (right and wrong) and to the left and right of 0.75 (low and high), did not reach significance (chi-square = 2.23, $df = 1$, NS; and chi-square = 2.75, $df = 1$, NS, for the data shown in Fig. 3a and 3b, respectively). This result contrasts with that of Lichtenstein *et al.* (1978) with both a student sample (their Fig. 2, page 555) and with a sample of the League of Women Voters (their Fig. 3, page 557), where respondents were found to be more likely to choose the wrong option when the true relative likelihood was low.

The rank-ordered correlations between assessed median marginals and true marginals was 0.24 ($p < 0.05$) for the students and 0.42 ($p < 0.01$) for the underwriters. Although the underwriters appeared better at judging the relative likelihood of pairs of events, the difference between correlations did not reach significance. It is interesting to note, however, that this trend reverses that shown in the direct comparisons.

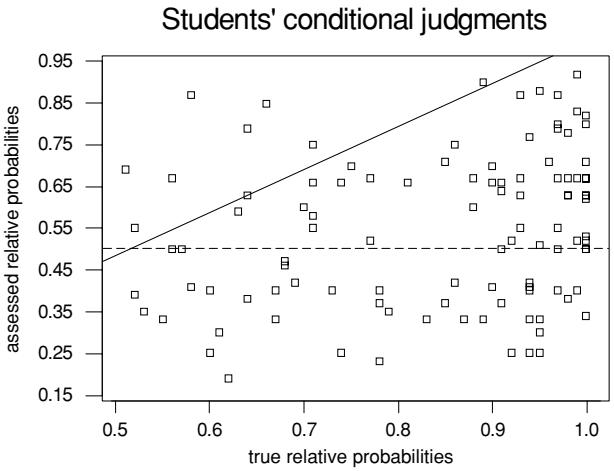

Fig. 4a. Students' pairwise assessments ($N = 105$) of "conditional" death rates each expressed as a probability that the truly most likely event of each pair will lead to death relative to the truly least likely event of each pair. Assessed relative probabilities are plotted against the true relative probabilities calculated in the same manner. Perfect assessments fall on the diagonal identity line and correct identifications of the truly most likely event of each pair fall above the dashed horizontal line.

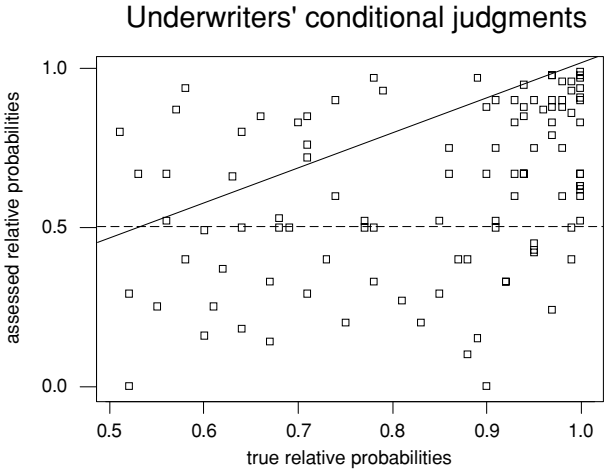

Fig. 4b. Underwriters' pairwise assessments ($N = 105$) of "conditional" death rates each expressed as a probability that the truly most likely event of each pair will lead to death relative to the truly least likely event of each pair. Assessed relative probabilities are plotted against the true relative probabilities calculated in the same manner. Perfect assessments fall on the diagonal identity line and correct identifications of the truly most likely event of each pair fall above the dashed horizontal line.

5.4. Paired Assessments: Group Conditional Analysis

Figs. 4a and 4b repeat the above analysis—this time for paired conditional assessments. Both graphs again evidence general underestimation of the true difference between pairs of events, this time in terms of their relative lethality. As previously, the tendency to choose the wrong option as more lethal appears to be associated with a smaller real difference in the

likelihood ratio between pairs. The rank-ordered correlations between assessments and true values were 0.25 ($p < 0.05$) and 0.42 ($p < 0.01$) for the students and the underwriters, respectively. Again, the underwriters appeared to be better, although the difference between these correlations was nonsignificant.

5.5. Paired Assessments: Individual Analysis

As outlined earlier, each respondent made either 28 or 21 paired marginal assessments within one block of life-threatening event items. This richness of data—at the level of the individual respondent—allowed us to measure an individual's error over his or her total number of assessments. For each respondent, we computed the mean average percentage error (MAPE), which is defined as:

$$\frac{\sum_{i=1}^{i=n} abs\left(\frac{a_i - t_i}{t_i}\right)}{n} * 100\%$$

where a_i = assessed relative probability of the truly most likely event of a pair occurring; n = the number of paired assessments made by an individual; and t_i = true relative probability of that event occurring.

The overall mean MAPEs for the marginal pairings were 42.7 (*SD* 11.3) and 36.4 (*SD* 8.2) for our students and underwriters, respectively. The means were significantly different ($t = 2.35$, $df = 73$, $p < 0.05$, two-tailed). Turning now to the conditional pairings, the overall mean MAPEs were 45.4 (*SD* 11.0) and 38.7 (*SD* 8.7) for our students and the underwriters, respectively. The difference between the means did reach significance ($t = 2.49$, $df = 73$, $p < 0.05$). Clearly, our underwriters were more accurate in their paired assessments than our undergraduate students, but only by a factor of, roughly, six percentage points of error.

Within our paired comparison data, we were also able to derive a measure of the reliability of an individual's assessments. Our analysis focused on the number of inconsistent triads within an individual's paired comparisons. For example, if an individual states that lethal event "a" is more likely than lethal event "b" and also that lethal event "b" is more likely than lethal event "c," he or she should not state that lethal event "c" is more likely than lethal event "a." Within the blocks of 28 paired comparisons, the maximum possible number of such inconsistent triads is 20. Within the block of 21 paired comparisons, the maximum possible number of inconsistent triads is 14. For the marginal pairwise comparisons, our students averaged 15% of inconsistent triads, whereas our underwriters averaged 18%. Nonparametric analysis revealed that this difference was not significant (Mann Whitney $u = 315$, $p > 0.05$, two-tailed). For the conditional pairwise comparisons, the average number of inconsistent triads was 8% for both samples. No significant correlations were obtained between individuals' degree of inconsistency and subsequent overall MAPE in either the marginal or conditional sections of our sets of paired comparisons, within our samples of students and underwriters.

5.6. Underwriting Experience and Paired Assessments

Our underwriters also provided individual biographical data on: the number of years that they had worked at the particular insurance company ($\bar{x} = 6.13$, $SD = 6.21$), the approximate number of years they had been doing life underwriting ($\bar{x} = 3.65$, $SD = 4.71$), the approximate number of hours they spent each week on underwriting ($\bar{x} = 11.74$, $SD = 12.33$), the approximate number of proposals they assessed per week ($\bar{x} = 54.4$, $SD = 43.7$), and the discretionary band within which they were able to make proposal evaluations without referral to a more senior underwriter. These bands were: (1) up to £30,000 with no additional evidence from that contained on the proposal form, (2) up to £50,000 with no evidence, (3) any proposal, no evidence, (4) up to £100,000 at ordinary rates but with additionally provided medical and other evidence on the risk, (5) up to £200,000 with evidence and the discretion to charge additional premiums, (6) up to £300,000 with evidence and discretion, and (7) any proposal within the company's limits. We coded these discretionary bands as 1 through 7. The frequency with which the underwriters fell into these bands was, 3, 3, 8, 5, 6, 9, and 3, respectively.

Our next analysis was to correlate our individual measures of underwriting experience with our individual measures of marginal/conditional and inconsistent triads. Our nonparametric correlation matrix is set out in Table III.

Several of the obtained significant correlations are unsurprising. For example, the higher the discretion of underwriting band that an individual is in, the longer he or she has worked at the insurance company, and the longer the number of years that he or she has been doing life underwriting. Some of the other obtained correlations are perhaps less obvious: the lower the individual's marginal MAPE, the greater

Table III. Correlation Matrix Between Measures of Underwriting Experience and Reliability/Validity of Both Marginal and Conditional Risk Assessments[1]

	YW	YE	HW	NP	DB	MM	CM	MIT	CIT
Years worked at Beta Co. (YW)	1.00								
Years experience life underwriting (YE)	0.76**	1.00							
Hours a week spent life underwriting (HW)	0.33	0.31	1.00						
Number of proposals assessed a week (NP)	0.13	0.11	0.51**	1.00					
Discretionary band (DB)	0.71**	0.87**	0.25	0.25	1.00				
Marginal MAPE (MM)	−0.16	−0.41**	0.35	0.21	−0.41**	1.00			
Conditional MAPE (CM)	−0.28	−0.27	−0.12	−0.36	−0.40**	0.38	1.00		
Marginal inconsistent triads (MIT)	0.27	0.18	0.40**	0.35	0.26	0.04	0.01	1.00	
Conditional inconsistent triads (CIT)	−0.27	−0.13	0.04	−0.06	−0.09	0.10	0.15	0.26	1.00

[1]Two asterisks next to obtained correlation indicate that the correlation was significant at the $p < 0.01$ level, two-tailed. Since we used an alpha of 0.01, 1 out of 100 of our obtained correlations can be expected to reach significance by chance alone.

the years doing life underwriting and the higher the discretionary band that this individual is in. Additionally, the higher the discretionary band that an individual is in, then the lower that individual's conditional MAPE. However, the degree of accuracy that an individual underwriter demonstrated on our two measures of marginal and conditional MAPEs was found to be unrelated, suggesting that performance on the marginal and conditional assessment tasks are underpinned by different cognitive processes. Our measures of inconsistency in triadic comparisons for both marginal and conditional assessments appear unrelated to underwriting experience—with the exception that the greater the number of hours a week that an underwriter spends underwriting, then the greater the number of marginal inconsistent triads evidenced in the responses of that underwriter.

Clearly, an individual's professional status within the life insurance company, as indicated by the discretionary band the individual is placed within, *is* indicative of the underwriter's performance on our paired-comparison assessment tasks. However, recall our earlier finding of only a small percentage difference in MAPEs between the performance of our samples of students and underwriters on these tasks.

6. DISCUSSION

Previous empirical research on expert-lay differences in risk judgment has been limited. Indeed, only one study (Slovic, Fischhoff, and Lichtenstein, 1985) has compared the *accuracy* of risk judgments of experts and lay persons (Rowe & Wright, 2001). This study based its conclusions—that experts are more accurate than various lay groups—on an expert sample comprising only 15 individuals from various professional disciplines performing estimations of risks with which they could not (in the most part) be expected to have any knowledge or experience. Additionally, the lay groupings in the Slovic *et al.* study may have been unaware that the evaluative standard for their "risk" estimates was that of the relative frequency of mortality. In the present study, we adopted the methodology of a second study, conducted by Lichtenstein *et al.* (1978), in order to explore in greater detail the nature of expert-lay differences in direct risk magnitude judgments. Furthermore, we took care to select a consistent sample of experts who regularly perform risk judgments in their professional lives (namely, insurance underwriters), and presented them with task items (i.e., the hazards to be assessed) that were ecologically valid.

The results from the study have revealed that although both lay and expert groups showed relatively good performance in terms of the ordering of the absolute likelihood (marginal) and lethality (conditional) 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 judgments, although less

so for conditionals. The student group was no worse at direct marginal or direct conditional estimation than the experts.[8]

Because the *direct* estimation of risks associated with potentially lethal events is an unusual task, even for our experts (at least for marginal estimates, although for conditional estimates the Chief Underwriter stated that this assessment mode captured the essence of her 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, analysis of MAPEs revealed that the experts *did* make significantly better judgments than lay persons 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, that is, in overly compressing their ranges of estimates.

So what do these results mean? The experts were, generally, a little better in their risk judgments than the lay persons, and the fact that expertise did make *some* difference was shown by the finding that the more senior underwriters made lower errors (in terms of MAPEs) in both the pairwise marginal and conditional tasks than those that were less senior. But the differences in performance between experts and lay persons were small in magnitude, and the nature of the biases (in terms of overestimating direct estimates and underestimating the differences in marginal/conditional riskiness between pairs of events), were common to both groups. These general biases seem to revolve around inadequate scaling of estimates, which is unsurprising, at least in the case of the lay group, given unfamiliarity with the raw figures related to risk. But why were the experts no better?

The previous evidence for experts being better at the judgments of risks (and, indeed, of perceiving risks in a different way than do lay persons) is not strong (see Rowe & 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. We attempted to ensure that our expert-task match was as strong as possible (given experimental limitations), and that ecological validity was high, and yet we 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 judgments. Indeed, in the training of underwriters, performance is assessed according to the similarity of junior underwriters' judgments to those of their seniors (Bolger *et al.*, 1989). Once "trained," underwriters receive infrequent performance-related, objective feedback about the correctness of their judgments and indeed it would be difficult to provide such feedback, given that a "poor" judgment 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/judgment feedback at all. From this, we suggest that expert-lay differences in the accuracy of such risk judgments, or in the nature of such judgments (given that the biases evidenced in this 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 judgments 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.

8 Expert-novice differences would be expected to be fairly large if they are to be at all practically relevant, in which case our sample size, although fairly small, should be adequate to achieve a reasonable level of power (0.8 or above, as suggested by Cohen, 1977). If this is the case, then, on the basis of our nonsignificant results, we can say our experts are mostly not doing any better than the novices (despite the confident responses of the Chief Underwriter to the contrary).

APPENDIX 1

Instructions

Each question in Part 1 consists of two different potentially lethal events. The question you are to answer is:

Which event is the most likely cause of death?

Suppose we randomly picked just one person from *all* the people now living in England and Wales. From which of the following events will that person be more likely to die?

A—a sporting accident
B—a stroke

Death due to each event is remotely possible. If you think that the person is MORE LIKELY to die from event A (sporting accident) than from event B (stroke) then you would circle "A." Alternatively, if you feel a stroke is a more likely cause of death than a sporting accident then you would circle "B."

Next we want you to decide *how many times more likely* this event is as a cause of death, as compared with the other event in the same question. For example, if you think that a sporting accident is twice as likely a cause of death than a stroke, then you would write a "2" in the space provided.

The pairs of potentially lethal events in Part 1 vary widely in their relative seriousness. For one pair, you may think that the two events are equally likely to cause death. If so, you should write the number "1" in the space provided for that pair. Or you may think that one event is 10 times, or 100 times, or even a million times as likely as the other event to cause death. For some pairs, you may believe that one event is just a little bit more likely to cause death than the other event. For this situation you will have to use a decimal point in your answer:

1.1 means that the more likely cause is 10% more likely than the other cause.
1.5 means 50% more likely, or half again as likely.
1.8 means 80% more likely.
2.5 means two and a half times as likely, etc.

Finally, before you start, we would like to draw your attention to the following notes, which may help you with your judgments. Please refer to these notes as you go along.

NOTES—Lethal Events Questionnaire

These notes may help you when making your judgments. ((The information in the double brackets following each item indicates the criterion we are using to determine whether an event happens to someone.))

Food poisoning:
Includes all types of poisoning due to organisms in food (salmonella, botulism, listeria, etc.).
Excludes accidental or deliberate poisoning by chemical agents introduced to foodstuffs.
((Notifications or hospital admissions))

Breast cancer:
With respect to women only.
((Notifications))

All accidents:
Includes any sort of accidental event.
Excludes disease and natural disasters (e.g., floods, landslides, etc.).
((Hospital admissions))

Pregnancy/childbirth:
Includes complications of pregnancy and childbirth occurring to women only.
Excludes infants, fetuses, etc.
((Reported cases))

Hodgkins disease:
This is a form of cancer affecting the lymph nodes.
((Hospital admissions))

Air accident:
Accidents occurring to inhabitants of England and Wales only (although accidents themselves may occur abroad).
Includes ALL mishaps involving airplanes either on ground, in air, or over/in sea. Acts of terrorism involving airplanes also included.
((Reported cases))

Bronchitis/emphysema:
Includes chronic obstructive airway diseases only.
Excludes asthma.
((Hospital admissions))

Suffocation/choking:
Either accidental or purposeful, self or other inflicted.
((Hospital admissions))

Infectious hepatitis:
Includes both hepatitis A and B.
Excludes chronic hepatitis and other related, non-infectious liver disease.
((Notifications))

Domestic accident:
Includes mishaps in and around the house.
Excludes diseases and violent acts.
((Hospital admissions))

Whooping cough:
Infectious disease.
((Notifications))

Lung cancer:
Includes cancer of the lungs only.
Excludes cancer of oesophagus, throat, etc.
((Notifications))

Diabetes:
Disorder of process by which the body uses sugars and other carbohydrates.
((Diagnoses))

Rail accident:
Includes train collisions only.
Excludes falls from train, person on track, etc.
((Reported cases))

Accidental poisoning:
Includes chemical substances taken unintentionally.
Excludes organic poisoning by foodstuffs and purposeful administration of harmful substances.
((Hospital admissions))

Alcohol/drugs:
Includes all substances purposely taken except as a means of suicide.
Excludes substances taken accidentally or for medical purposes.
((Hospital and special unit admissions))

Measles:
Infectious disease.
((Notifications))

Fire and flames:
Includes burns and smoke inhalation only.
((Hospital admissions))

Asthma:
Includes both allergic and late-onset asthmas.
Excludes other obstructive airway diseases such as bronchitis and emphysema.
((Hospital admissions))

Polio:
Infectious disease of the nervous system.
((Notifications))

Stomach cancer:
Includes cancer of alimentary tract, oesophagus, and intestines, as well as stomach.
((Notifications))

Car accident:
Includes passengers and drivers of any motor vehicle (cars, buses, motorcycles, lorries, etc.).
Excludes cyclists and pedestrians.
((Reported severities))

Smoking related:
Includes deaths due to lung cancer, heart disease, and respiratory diseases attributable to active smoking only.
((Estimated proportion of admissions/notifications of above three ailments))

Salmonella:
Bacterial food poisoning.
((Reported cases))

Leukaemia:
Acute and chronic cancers characterized by an abnormal increase in the number of white blood cells.
Excludes other, noncancerous blood diseases.
((Notification))

Lightning:
Act of God!
((Reported cases))

Hypertension:
Includes high blood pressure due to underlying pathology (e.g., kidney disease, hardening of the arteries, etc.).
Excludes high blood pressure with no known cause and temporarily elevated blood pressure due to, e.g., exertion.
((Hospital admissions))

Meningitis:
Infectious disease of the nervous system.
((Notifications))

Road accident:
Includes pedestrians and cyclists hit by motor vehicles.
Excludes drivers and passengers of motor vehicles.
((Reported severities))

Benign tumors:
Includes tumors classified as "benign" due to location and/or noninvasive nature (i.e., nonspreading).
Excludes tumors classified as "malignant" and benign tumors later reclassified as malignant.
((Notifications))

Electric shock:

Shock from mains voltages or higher.

((Reported cases))

APPENDIX 2

Instructions

In Part 1 we asked you to judge the relative likelihood of death from pairs of events.

In Part 2 we would like you to judge the same pairs of events BUT this time we would like you to answer:

Which event is the most likely to cause death given it happens to someone?

Suppose we randomly picked just one person from *all* the people now living in England and Wales. Let's say, for example, that *one* of the two following potentially lethal events happens to this person:

A—he or she is involved in a motorcycle accident
B—he or she is diagnosed as having influenza

If you think the person is MORE LIKELY to die as a result of event A (motorcycle accident) than as a consequence of event B (influenza) then you would circle "A." Alternatively, if you feel influenza has more serious consequences in terms of mortality than a motorcycle accident, then you would circle "B."

Next we want you to decide *how many times more likely* this event is to cause death, as compared with the other event given in the same question. For example, if you think flu is twice as likely to cause death to someone than him or her being involved in a motorcycle accident, then you would write a "2" in the space provided.

Please mark your judgments in a similar manner to the way you did in Part 1 (i.e., "1" means equally likely, "1.5" means 50% more likely, "2" means twice as likely, "100" means 100 times more likely, "1,000,000" a million times more likely, etc.).

Again you may refer to the notes relating to each event that are given in Part 1.

APPENDIX 3

Instructions

What you have been judging in Parts 1 and 2 are what are known as *marginal* and *conditional* probabilities of death due to various potentially lethal events.

The *marginal* probability in this case refers to the number of people dying in England and Wales each year as the result of a particular lethal event.

For example, in 1986 in England and Wales 163,200 died as a result of heart disease (excluding stroke).

This is normally expressed as a mortality rate, e.g., 326/100,000 for the heart disease example above.

Thus 326 people out of every 100,000 in the entire population of England and Wales died as a result of heart disease in 1986.

The *conditional* probability in this case refers to the likelihood of someone dying in England and Wales each year given that a particular event has happened to them.

For example, in 1986 in England and Wales 348,718 people were admitted to hospital with diagnosed heart disease (excluding stroke).

As we know from the marginal probability above, 163,200 people actually died of heart disease that year. Assuming (as appears to be the case) that the number of fatalities and hospital admissions remains fairly constant over years for heart disease, we can *estimate* the likelihood of death from heart disease for someone diagnosed as suffering from this ailment.

This means that just under half the people who suffer from heart disease subsequently die from it. If the conditional probability had been 1 this would mean that *all* people suffering from heart disease die as a result of it. If the conditional probability had been 0 this would mean that *no* people suffering from heart disease die as a consequence of that ailment.

It has been found that people give very different estimates of marginal and conditional probabilities for the same events, depending on how these probabilities for the same events are asked for. Generally, people seem to find it easiest to give estimates in the form of the paired comparisons you made in Parts 1 and 2. However, we would also like you to try and give direct estimates of the marginal and conditional values for each of the lethal events you have previously considered.

REFERENCES

Barke, R. P., & Jenkins-Smith, H. C. (1993). Politics and scientific expertise: Scientists, risk perception, and nuclear waste policy. *Risk Analysis, 13*(4), 425–439.

Bolger, F., & Wright, G. (1994). Assessing the quality of expert judgment: Issues and analysis. *Decision Support Systems, 11*, 1–24.

Bolger, F., Wright, G., Rowe, G., Gammack, J., & Wood, B. (1989). LUST for life: Developing expert systems for life assurance underwriting. In N. Shadbolt (Ed.), *Research and development in expert systems VI*. Cambridge: CUP.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

Einhorn, H. J. (1980). Learning from experience and suboptimal rules in decision making. In T. Wallsten (Ed.), *Cognitive processes in choice and decision* (pp. 1–20). Hillsdale, NJ: Erlbaum.

Flynn, J., Slovic, P., & Mertz, C. K. (1993). Decidedly different: Expert and public views of risks from a radioactive waste repository. *Risk Analysis, 13*(6), 643–648.

Gutteling, J. M., & Kuttschreuter, M. (1999). The millennium bug controversy in the Netherlands? Experts' views versus public perception. In L. H. J. Goossens (Ed.), *Proceedings of the 9th annual conference of risk analysis: Facing the millennium* (pp. 489–493). Delft: Delft University Press.

Keren, G. (1987). Facing uncertainty in the game of bridge: A calibration study. *Organizational Behavior and Human Decision Processes, 39*, 98–114.

Kraus, N., Malmfors, T., & Slovic, P. (1992). Intuitive toxicology: Expert and lay judgments of chemical risks. *Risk Analysis, 12*(2), 215–232.

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

Maxwell, R. J. (1999). The British government's handling of risk: Some reflections on the BSE/CJD crisis. In P. Bennett & K. Calman (Eds.), *Communications and public health* (pp. 94–107). Oxford: Oxford University Press.

McDaniels, T. L., Axelrod, L. J., Cavanagh, N. S., & Slovic, P. (1997). Perception of ecological risk to water environments. *Risk Analysis, 17*(3), 341–352.

Mumpower, J. L., Livingston, S., & Lee, T. J. (1987). Expert judgments of political riskiness. *Journal of Forecasting, 6*, 51–65.

Murphy, A. H., & Brown, B. G. (1985). A comparative evaluation of objective and subjective weather forecasts in the United States. In G. Wright (Ed.), *Behavioural decision making*. New York: Plenum.

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

Rowe, G., & Wright, G. (1993). Expert systems in insurance: A review and analysis. *International Journal of Intelligent Systems in Accounting, Finance, and Management, 2*, 129–145.

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

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., & Lichtenstein, S. (1985). Characterizing perceived risk. In R. W. Kates, C. Hohenemser, & J. X. Kasperson (Eds.), *Perilous progress: Managing the hazards of technology* (pp. 91–125). Boulder, CO: Westview.

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

Wright, G., Pearman, A., & 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.