ORGANIZATIONAL BEHAVIOR AND HUMAN DECISION PROCESSES
Vol. 70, No. 2, May, pp. 117–133, 1997
ARTICLE NO. OB972697

Taking Advice: Accepting Help, Improving Judgment, and Sharing Responsibility

Nigel Harvey and Ilan Fischer

University College London, United Kingdom

Why do people take advice? To find out, we provided a low, medium, or high level of training on a task in which judgments varied in importance. Then, in a test session, we eliminated feedback but made advice available. Different features of our results suggest that there are three distinct reasons for taking advice. First, even experienced judges took some advice from novices: people appear reluctant to reject completely the help offered to them. Second, all judges took more advice from advisors more experienced than themselves, and the amount that they took was related to the difference in level of experience: people appeared to be trying to use advice to improve their judgments. Third, experience enabled people to distinguish judgments on the basis of their importance. Experienced judges took about twice as much advice for the most important judgments: they appeared to be sharing responsibility when the risk associated with error was high. Thus, we model advice-taking in terms of three components: accepting help, improving judgment, and sharing responsibility.

In many institutions, judges or executives take individual decisions after conferring with one or more advisors or consultants. With a few rare exceptions (e.g., Gardner & Berry, 1995; Sniezek & Buckley, 1993, 1995), judge-advisor systems of this type have not been studied systematically by psychologists. To investigate them, a person can be asked to make an initial judgment on the basis of some evidence, provided with advice from some source, and then asked for a final judgment. Influence of the advice is measured by the size of the shift between the initial and final judgment. Effect of this influence is measured by the improvement in the quality of the judgment. Here we shall be concerned primarily with factors that determine the influence that advice has on a judge.

Hogarth and Einhorn's (1992) Belief Adjustment model provides some insight into how advice may be integrated with an individual's prior beliefs. Consider a doctor who has assessed a terminally ill patient's life expectancy and then asked a colleague for an independent opinion. According to Hogarth and Einhorn (1992), the doctor forms a final opinion by taking a weighted average of his or her prior opinion and the colleague's advice. The weight given to the advice depends on the person's sensitivity to new information. This, in turn, depends on situational variables such as the credibility of the source of the new information (e.g., the colleague's seniority). However, Hogarth and Einhorn (1992) also argue that sensitivity to information varies with personal variables: their example of this is a scientist who is not sensitive to information about a favorite theory. Their model therefore implies that people receiving advice should take their beliefs about their own level of expertise into account as well as their beliefs about the expertise of their advisors.

At present, there appears to be no direct evidence that either of these factors is important in determining the influence of advice on judgments themselves. However, recent work on people's confidence in their judgments suggests that they take their perceptions of both their own expertise (Arkes, Christensen, Lai, & Blumer, 1987; Trafimow & Sniezek, 1994) and their advisors' expertise (Sniezek & Buckley, 1995) into account. Arkes *et al.* (1987) and Trafimow and Sniezek (1994) found that experimental manipulations designed to reduce participants' perceptions of their expertise in answering a set of questions decreased their confidence in their answers, whereas those designed to elevate their perception of their expertise increased their confidence in their answers to the same set of questions. Sniezek and Buckley (1995) examined whether two advisors'

Parts of this paper were presented at the Seventh Meeting of the European Society for Cognitive Psychology, Lisbon, 1994, and at the Sixteenth International Symposium on Forecasting, Istanbul, 1996. The research was in part funded by Economic and Social Research Council Grant R000236827.

Address correspondence and reprint requests to N. Harvey, Department of Psychology, University College London, Gower Street, London WC1E 6BT, UK. Fax: +44 171 436 4276. E-mail: n.harvey@ucl.ac.uk.

0749-5978/97 $25.00

confidence in their recommendations influenced a judge's selections in a two-alternative question-answering task. Both confidence and performance tend to increase with practice. Consequently, judges may regard confidence as signifying expertise and, therefore, be more likely to take advice from confident individuals. Results confirmed this: when the two advisors disagreed, judges tended to accept the recommendations from the more confident one.

Research in two other areas also suggests that people will take both their own and their advisors' expertise into account. First, Birnbaum, Wong, and Wong (1976) and Birnbaum and Stegner (1979) showed that people combining information from different external sources take the perceived expertise of those sources into account: if people presented with advice regard themselves as well as their advisors as sources of information, the way they combine external information (advice) with internal information (their own judgment) should be influenced by the levels of expertise that they perceive to be present both in themselves and in their advisors.

Second, work on attitude change has shown that the effectiveness of a persuasive message depends on the expertise attributed to its source (e.g., Petty, Cacciopo, & Goldman, 1981).[1] However, reliance on peripheral cues such as this tends to give way to analysis of message content when recipients regard themselves as knowledgeable or experienced in dealing with the issue to which the message refers (e.g., Wood, 1982; Wood, Kallgren, & Priesler, 1985) or when the issue is taken to be a relatively important one (e.g., Chaiken & Maheswaran, 1994). People are more likely to carry out mental work when they perceive it to be easy (because they are practised at it) or when they need to be accurate (because the outcome is important) (cf. Payne, Bettman, & Johnson, 1993).

To what extent can these contingent processing notions be generalized beyond persuasion to advice-taking? Given the work discussed earlier (e.g., Sniezek & Buckley, 1995), it is reasonable to expect that novice judges will be more likely than experienced judges to take advice and that they will be more influenced by the level of expertise attributed to their advisors. But what about the effect of the importance of the judgment? There are a number of possibilities: we shall briefly outline two of them.

First, novices may, rightly or wrongly, see themselves as capable of making reasonably good judgments. However, they perceive good judgment to be so cognitively demanding that they normally decide to place heavy reliance on their advisors' views. Only when judgments are particularly important do they decide to carry out the additional mental work needed to avoid this reliance (cf. Chaiken & Maheswaran, 1994). This suggests that novices should place *less* reliance on advisors and be *less* influenced by advisors' levels of expertise when judgments are important.

Alternatively, novices may not see themselves as capable of making particularly good judgments. However, they normally rely primarily on their own opinions rather than on those of their advisors because what they find cognitively demanding is not so much making the judgments in the first place as integrating them with the views of their advisors. Only when judgments are particularly important do they decide to carry out the additional mental work needed to perform this integration. This suggests that novices should place *more* reliance on advisors and be *more* influenced by advisors' levels of expertise when judgments are important.

EXPERIMENT 1

On the basis of these considerations, we designed a first experiment to answer two questions. Are novice judges more influenced by advisors who have greater experience at the task? How is the extent to which they are influenced by advice affected by the importance of the judgment that they have to make?

The Task

We employed a cue-learning task. This approach enabled us to study advice-taking after participants had acquired different amounts of experience at the task. It appears that people use information about the amount of experience acquired in a task as a means of judging expertise in performing it. This is suggested by the illusions of learning that occur when people expect practice to be effective but it is not (e.g., Adams & Goetz, 1973; Harvey, 1994; Marteau, Johnston, Wynne, & Evans, 1989). However, it remains possible, if unlikely, that people have direct access to their own trace strengths in cue-learning tasks and can exploit that information. Hence, we designed our task with the aim of ensuring that experience really produced learning and then checked that it did so.

A simple algorithm was used on each trial to generate a criterion value from cue values. Participants were not given the algorithm but had to predict the criterion

[1] As far as we are aware, the connection between the influence of persuasion on attitudes and the influence of advice on judgment has not been made explicit before. The parallels seem clear to us but we shall use them with caution. Attitudes and judgments differ in generality of their referents, and persuasion and advice differ in intent. Nevertheless, it remains possible that the basic mechanisms underlying the effects of influence are fundamentally the same in the two cases. Direct empirical comparisons would be needed to ascertain whether this is so.

value that it would produce on the basis of the cue values. To achieve a reasonably rapid learning rate, we used a simple task with just two cue values and no error component (cf. Hammond & Summers, 1965; Summers & Hammond, 1968). We also presented cues pictorially rather than numerically. By providing clarity at the expense of precision (Cooksey, 1996; Stewart, 1988), this can be expected to speed up learning but to limit its asymptotic level. Finally, we provided participants with a cover story to frame their judgments. Scenarios do not always facilitate performance (Sanderson, 1989). However, they tend to do so when they embed the task in a setting known to the participants (Berry & Broadbent, 1984; Stanley, Mathews, Buss, & Kotler-Cope, 1989) and when they generate more accurate beliefs about the formal relationship holding between cue and criterion values (Adelman, 1981; Sanderson, 1989).

During the training phase, participants were given outcome feedback after each trial that specified the correct criterion value. With simple cue-learning tasks, outcome feedback is effective in producing learning (e.g. Schmitt, Coyle, & Saari, 1977). However, we used a multiplicative rather than an additive relationship between the two cues. While this makes tasks simpler (i.e., easier to learn) in some situations, it makes them more complex in others (Edgell, 1993). We therefore carried out pilot studies to check that reasonably rapid learning occurred in the task. These confirmed that it did so. After just 30 trials, performance was significantly improved but still below what could be achieved with more practice. We shall characterize people after this level of training as *novices*. After 100 trials, performance started to asymptote: we shall characterize people at this stage of learning as *semi-experienced*. Finally, after 240 trials, the performance asymptote was clearly evident: we shall characterize people who have had this amount of practice as *experienced*. The current experiment investigated advice-taking in novices. In later experiments, we studied it in semi-experienced and experienced judges.

The training phase was followed by a test session in which no feedback about the correct criterion value was given. On each test trial, participants first gave their initial estimate of the criterion value. Next, they received a recommendation from the advisor. After that, they produced a final estimate based both on the cue values of the stimulus in front of them and on the advice that they had received. We estimated willingness to take advice by measuring the change between the pre-advice and post-advice criterion estimates.

People were led to believe that different advisors made recommendations to them. A third of their advisors had 30 training trials, a third had 100 training trials, and a third had 240 training trials. When they were told an advisor's recommendation, they were also told how many training trials that particular advisor had received. As we mentioned above, we expected people to perceive advisors who had received more training as having greater expertise in the task.

In fact, all advice was generated by the experimenters, and it did not depend in any way on the number of practice trials that were specified for the advisor. Impressions of expertise were under our control. This was important for two reasons. First, it ensured that variability arising from individual differences in advisors' learning rates did not contaminate our results. Second, it allowed us to produce systematic variation in the quality of the advice that was independent of the level of training attributed to the advisors. One third of the recommendations provided by advisors with each level of training specified the correct criterion value; one third underestimated that value by a given amount; the remaining third overestimated it by the same amount. Thus any effect of the expertise attributed to the advisors cannot arise from differences in the actual quality of advice they provide.[2]

To manipulate the importance of judgments, we used a scenario in which participants made forecasts to provide farmers with immediate compensation from the government for loss of cattle from outbreaks of disease. Specifically, they had to predict how many cattle would die on the basis of the size of the land area affected and the type of disease. Error in forecasts would lead to inappropriate compensation which would incur the wrath of either the farmers or the government when the outbreak had finished and the true number of deaths had been ascertained.

We expected participants to regard more severe outbreaks as more important because they provided the potential for more costly forecast errors. In an attempt to reinforce the salience of this link, participants were given points for their performance during their training. A correct judgment warranted 100 points. An incorrect one received points equal to 100 divided by the absolute difference between judgment and criterion. Hence the number of points they received after a given percentage error in their judgment was inversely related to the severity of the outbreak. For example, when the criterion was 20, a 10% judgment error still resulted

[2] Of course, if people were able to assess quality of advice, they would be able to determine that advice from their experienced advisors was no better than that from other advisors. However, this would result in them being no more prone to take advice from experienced advisors than from other advisors: it would eliminate the predicted effect. In fact, the analyses that we report demonstrate that judges were no more influenced by correct than by incorrect advice: they could not assess advice quality.

in an award of 50 points but, when the criterion was 50, it produced only 20 points.

Method

Participants. Thirty undergraduate students from the psychology department at University College London acted as participants. The experiment took each of them about 60 min to complete.

Stimuli. Stimuli were computer-controlled and presented on a color monitor. They comprised colored circles presented for approximately half a second in a gray frame. The frame measured 18 cm by 11.5 cm. The size and color of the circles were cues for the participants' responses, and so they varied from trial to trial. Position of the circles within the frame was also varied from trial to trial to ensure that the distance between the edge of the circle and the frame could not act as a substitute for the size cue.

A request for participants to make an estimate of the criterion value appropriate to the size and color of the displayed circle appeared below the frame. Participants could see their response as they typed it in. During training, feedback information specifying the correct criterion value and the number of points for their judgment of it appeared above the frame for 3 s after the response had been entered.

Criterion values (Y) were related to the area (X) and color of the displayed circle by the following algorithm:

$$Y = \alpha\beta X,$$

where

$$X = \pi r^2$$

Here α was a constant (0.001), r was the radius of the circle, and β took on a value of one half when the circle was blue, one when it was purple, two when it was green, and three when it was red.

Within our scenario, color represented the type of virus causing disease, X was the area affected by it, and Y was the number of cattle dying from it.

Design. On each of the 30 training trials, the circle that was presented was determined by randomly selecting a radius from a normal distribution (mean: 70 screen pixels; standard deviation: 17 screen pixels) and by randomly choosing one of the four colors with a probability of 0.25 for each one.[3] Participants made a single judgment when shown the stimulus and then received feedback information.

After training, participants completed 72 test trials. On each one, they first made their initial estimate of the criterion value in the same way as they had done during training. However, once they had done this, the display changed. In the upper half of the screen, they saw two boxes: the left one repeated the initial estimate they had given, and the right one specified the number of training trials they had received. In the lower half of the screen were two more boxes for corresponding details of the advisor's estimate and training level. These boxes flashed for a short but variable period: participants were told that the computer was searching for the information relevant to the particular outbreak under consideration. After details of the advisor's estimate and training level had been displayed, participants typed in their final estimate.

Correct advice (viz. advice giving participants the correct answer) was just the criterion value (Y) that the algorithm produced from the cue values (r, β) selected on that trial. To produce incorrect advice (viz. advice giving participants the incorrect answer), one standard deviation (17 pixels) was added to or subtracted from the value for the radius (r) that had been selected on that trial. This new r value was then inserted into the algorithm to obtain a criterion value greater or less than the correct one.

The training level of the advisor was given as 30 trials, 100 trials or 240 trials. Each of these three levels of advisors' training was specified on a third of the 72 test trials. Within each level, a third of the advice was correct, a third was greater than the correct value, and a third was less than it. Thus there were eight test trials for each of the nine combinations of advisors' training level and advice quality. Outbreak severity was varied over these eight trials by using different combinations of circle radius and color for each one. There was one high area outbreak ($r > 70$ pixels) and one low area outbreak ($r < 70$ pixels) for each of the four diseases. This gave the following criterion numbers of cattle deaths (with their associated cue values in parentheses): 4 (35 pixels, purple); 7 (65 pixels, blue); 14 (95 pixels, blue); 18 (75 pixels; purple); 19 (55 pixels; green); 19 (45 pixels, red); 45 (85 pixels, green); 104 (105 pixels, red).

Procedure. Participants were run individually in experimental cubicles. They were told that land in Britain can be infected with various different cattle viruses. More cattle die when the outbreak covers a larger amount of land and when the virus is more dangerous. When an infection is reported, the Ministry of Agriculture makes an inspection to determine how serious the

[3] Thus the distribution of r^2 was positively skewed. Such distributions provide good representations of accident event magnitudes and hence appear appropriate for describing the severity of outbreaks of each disease within our scenario.

situation is. This is because they have to make a forecast of how many cattle will die. These forecasts are needed because initial compensation is paid on the basis of them. This ensures that farmers have some immediate recompense rather than having to wait until the outbreak has finished before making a claim.

Forecasts need to be accurate so that farmers receive neither too little compensation (unfair to them) or too much (unfair to the taxpayer). Consequently, the Ministry of Agriculture trains their inspectors to provide them with some skill in forecasting the severity of disease outbreaks. Historical data are used so that trainees can be told the actual number of cattle deaths in an outbreak after they have made each forecast. To motivate their learning, they are rewarded with points. These are on a sliding scale: the greater the difference between the forecast and the actual number of cattle deaths, the lower the number of points.

After the scenario had been set, participants were given the following specific instructions.

> In the training session, you will see brief presentations of coloured circles. The size of a circle reflects the amount of land that is infected. The color of the circle gives you information about how dangerous the virus is. Red signals the most dangerous virus; green the next most dangerous; purple the next most dangerous; and blue the least dangerous. After you have seen the circle, type in your forecast of the number of cattle that you think will die from the outbreak. After you have done this, you will be told how many actually did die according to the Ministry's historical records. You will be able to compare this with your forecast and thereby improve your future forecasts. After 30 assessments, your ability to make these judgments should have improved considerably.
>
> In the second part of the experiment, you will use your training to make forecasts of how many cattle will die in current outbreaks. Obviously you can receive no information about the accuracy of these assessments because cattle are still dying. However, after making your assessment, you will be told the views of someone who has also inspected the same outbreak. You will also be told how much training this person received from the Ministry. After getting this information you will be given the opportunity of revising your original forecast. Do not feel obliged to make use of this information. It is up to you whether you take it into account.

After reading these instructions, participants completed the experiment. When they had finished, they were thanked for their participation and debriefed. There was no indication that they had been aware that their advisors had been notional rather than actual or that they had realized that the quality of the advice received from these advisors had been independent of the level of expertise attributed to them.

Measures of forecast accuracy and advice-taking. - Selecting measures of forecast accuracy and advice-taking is not a simple matter. All indices have advantages and disadvantages, and the value of each of them relative to others varies from situation to situation.

Various measures of overall forecast accuracy have been proposed (Makridakis, Wheelwright, & McGee, 1983; Wheelwright & Makridakis, 1985). Some are relevant only to time-series forecasting, but three can be considered as candidates for measuring forecasting accuracy in our task. The first of these is the root mean squared error (RMSE). This is calculated by taking the square root of the average of the squared differences between forecasts and criterion values. The second is the mean absolute percentage error (MAPE). This is the mean of the absolute values of the percentage errors in the forecasts. (Percentage error is the difference between the forecast and criterion expressed as a percentage of the criterion.) The third is the median absolute percentage error (MdAPE). This is the median of the absolute values of the percentage errors in the forecasts.

Carbone and Armstrong's (1982) survey of 145 forecasting experts showed that up to the early 1980s RMSE was the preferred accuracy measure among them. It has a number of advantages, not least of which is that it (or derivatives of it) can be partitioned into psychologically meaningful components (Harvey & Bolger, 1996; Lee & Yates, 1992; Stewart & Lusk, 1994). However, it is not unit-free. (Thus it is more appropriate for comparing the ability of different forecasting methods—or forecasters—to forecast the same type of data than for comparing the ability of the same or different forecasting methods to forecast different types of data.)

As the need for unit-free measures became recognized (e.g., Chatfield, 1988), forecasting experts came to prefer MAPE over RMSE (Ahlburg, 1992; Armstrong & Collopy, 1992). However, two problems associated with use of MAPE need to be borne in mind (Armstrong & Collopy, 1992; Fildes, 1992). First, it is appropriate only for ratio-scaled data (i.e., data with a meaningful zero). Second, it is bounded on the low side but it is unbounded on the high side. This bias in favor of low forecasts needs to be taken into account when MAPE scores are analyzed by applying an appropriate transformation to them (Box & Cox, 1964).

In selecting an error measure, it is important to take reliability and validity into consideration. Armstrong and Collopy (1992) assessed reliability by studying the extent to which an error measure produced the same accuracy rankings for 11 forecasting methods when it was applied to five different subsamples of 19 data sets. For each error measure, they calculated Spearman rank-order correlation coefficients for the accuracy rankings between each pair of subsamples and then averaged the resulting 10 pairwise correlations to obtain an overall estimate of reliability. Estimates for RMSE, MdAPE, and MAPE turned out to be 0.2, 0.4,

and 0.5, respectively. Furthermore, MAPE had the highest construct validity ($r_s = 0.9$).

To maximize reliability and validity, we followed the practice currently preferred by forecasters and adopted MAPE as our accuracy measure. Our data have a meaningful zero (i.e., no dead cattle), and so, in this respect, it is an appropriate index. However, deciding to use MAPE does not constitute an unconditional endorsement of this measure: as we mentioned above, it is important to bear in mind the problems associated with it when carrying out analyses. In particular, data transformation may be appropriate (Box & Cox, 1964).

Our advice-taking paradigm is one of a number concerned with the influence that an external estimate or expression of the criterion value has on quantitative judgment. Percentage shift is usually taken as a measure of this influence. For example, in their study of the hindsight bias, Hell, Gigerenzer, Gauggel, Mall, and Müller (1988) first required people to judge various quantities, later presented them with the correct values and asked them to use their memory to decide whether each of their responses had differed from the correct one by more than a factor of two, and finally told them to reproduce their original responses. To measure the effect that processing the correct values had on memory for responses, they used the percentage shift measure: the difference between the original and final response was expressed as a percentage of the difference between the original and correct response.

We adopted this same percentage shift measure: the influence of advice was estimated as the difference between the original and final judgment expressed as a percentage of the difference between the original judgment and the advice. Thus someone whose original judgment is 90, who receives advice of 100, and who then moves to 95 shows a 50% shift. If they had perversely moved to 85 after the same advice, they would have shown a shift of −50%; if they had moved to 105, they would have shown a shift of 150%. In other words, the measure is unbounded on both its high and low sides. (In fact, in our study virtually all shifts were between 0 and 100%.)

Of course, as with all percentage measures, it is important to recognize that shifts that differ in absolute terms may be characterized as equivalent. For example, someone who moves from 90 to 95 after advice of 100 shifts the same 50% as someone else who moves from 60 to 80 after advice of 100.

Results

Trials in which participants gave criterion values that were more than five times the correct value were excluded from the analysis. This filter served to eliminate responses that were clear mistypings: for example, one participant gave 60 as an initial estimate, received advice that it should be 55 and then typed 5959 instead, presumably, of 59. This procedure resulted in less than one per cent of the data being excluded.

In this section, we shall briefly comment of the effectiveness of training and then present analyses of the level of advice-taking and of the effect of advice on judgment quality.

Figure 1a shows that MAPE scores decreased over the three blocks of the training session. Slope of the graph of the log of their standard deviations against the log of their means was 0.08, indicating that data transformation was unnecessary (Box & Cox, 1964). Comparison of the first and last blocks indicated that learning occurred ($t(29) = 2.75$; $p < .01$). However, as we shall see when comparing these results with those of later experiments, it did not reach its asymptotic level.

Figure 1b shows the mean values of percentage shifts in judgment as a function of the level of training attributed to the advisors and the outbreak severity. Inspection of these data suggests that shifts increased with the level of training attributed to advisors but were unaffected by outbreak severity. To confirm this impression and to check whether people were sensitive to advice quality, we carried out an analysis of variance using advisors' training level, outbreak severity, and advice correctness as within-subject variables. Effect of advisors' training level was highly significant ($F(2,2050) = 23.90$; $p < .0001$) but neither of the other main effects nor any of the interactions approached significance. There was no evidence that people were sensitive either to outbreak severity or to advice correctness.

Figure 1c shows the reduction in MAPE between judges' initial and final judgment as a function of the training attributed to the advisors and the correctness of advice they gave. To study the effect of advice, we carried out an analysis of variance on the MAPE scores using advisors' training level, outbreak severity, advice correctness, and occasion (pre- vs post-advice) as within-subject variables. However, as the graph of the log of the cell standard deviations against the log of their means had a slope of .44, we first subjected them to a square root transformation (Box & Cox, 1964).

The effect of occasion was highly significant ($F(1,4184) = 44.95$; $p < .0001$), demonstrating that advice improved judgments. Initial judgments were so bad that even the incorrect advice that we gave produced final judgments that were an improvement on them (Fig. 1c). However, greater improvements were produced by correct advice, an unsurprising phenomenon that led to a significant main effect of advice correctness ($F(1,4184) = 30.84$; $p < .0001$) and a significant interaction between occasion and advice correctness ($F(1,4184) = 40.41$; $p < .0001$).

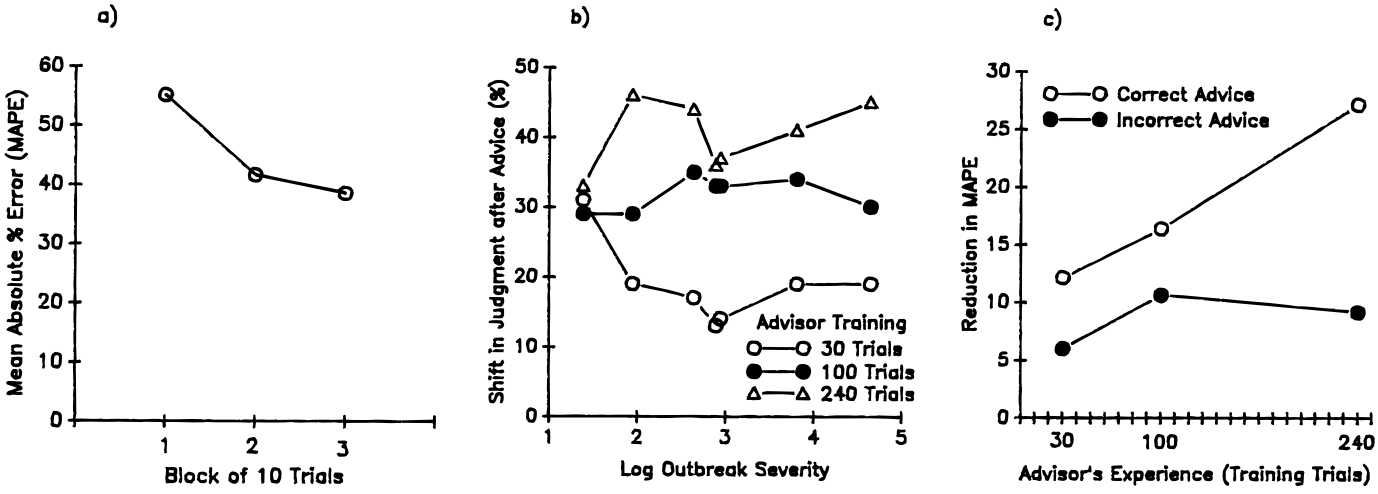

FIG. 1. Experiment 1. (a) Mean absolute percentage error (MAPE) over three blocks of 10 trials in the training session. (b) Percentage shift in judgment after advice. Use of log scale on abscissa is to aid visual clarity. (c) Reduction in MAPE after advice.

An interaction between occasion and advisors' training level ($F(2,4184) = 3.24$; $p < .05$) showed that improvements were greater when advisors were attributed with more expertise. Again, this is not surprising: judges took more advice from more experienced advisors (Fig. 1b), and, therefore, its beneficial effect was greater. One might expect that taking more advice would magnify the difference between the benefits of taking good advice and poor advice. Figure 1c suggests this is so. However, the three-way interaction between occasion, advice correctness, and advisors' training level did not quite reach a conventional level of significance ($F(2,4184) = 2.77$; $p = .06$); only the two-way interaction between the latter two variables did so ($F(2,4184) = 4.30$; $p < .05$).

For the six most severe outbreaks, pre-advice and post-advice MAPE scores were unaffected by outbreak size: they averaged 35 and 28%, respectively. However, for the two least severe outbreaks, MAPE scores were much higher, were reduced more by advice, and were affected more by advice correctness. These differences produced a main effect of severity ($F(7,4184) = 98.88$; $p < .0001$) and interactions between this factor and occasion ($F(7,4184) = 3.66$; $p < .001$) and between it and advice correctness ($F(7,4184) = 3.38$; $p < .01$). These effects appear to be artifacts of the small absolute size of the criterion values in the two least severe outbreaks and of the experimental requirement to produce judgments as whole numbers. For example, initially judging that the least severe outbreak will kill six cattle and revising this to five after advice gives pre- and post-advice MAPE scores of 50 and 25%, respectively; changing the judgments by a single unit produces a much larger MAPE reduction than the 7% average observed for the six most severe outbreaks.

Discussion

Our novice judges took more advice from advisors whom they regarded as more highly trained. Thus one of the generalizations derived from work on the effectiveness of persuasive messages (e.g., Petty & Cacciopo, 1986) applies equally to advice-taking: non-experts are influenced by information that they interpret as diagnostic of the credibility or expertise of the source of a message. In Sniezek and Buckley's (1995) experiment, this information specified level of confidence; in our experiment, it specified level of training.

Our predictions about effects of outbreak severity fared less well. We assumed that people would treat judgments about more severe outbreaks as more important because the consequences of forecast error were greater. On the face of it, this does not seem unreasonable: a 35% forecast error for an outbreak killing 14 cattle corresponds to a mean over- or underestimation of five cattle, whereas the same error for an outbreak killing 104 cattle corresponds to a mean over- or underestimation of 36 cattle. The average error in compensation would be more than seven times greater in the latter case.

We considered two possibilities. First, novices may be able to improve their own judgments by putting more cognitive effort into them. If they could, they would be more likely to do so when they perceive their task as important (e.g., Payne *et al.*, 1993). Hence, they should place *less* reliance on advice when outbreaks are severe. Alternatively, people may be disinclined to make careful and discriminating use of advice because they find doing so cognitively demanding. If this is so and if they see advice as potentially beneficial, they would be more likely to make use of it when they perceive their task

as important (Payne *et al.*, 1993). Hence, they should place *more* reliance on advice when outbreaks are severe.

As we have seen, neither of these patterns was present in the data. This may have been because outbreak severity was insufficiently salient as an indicator of task importance. We hoped that the scenario itself would be enough to establish this relationship but, to reinforce it, we gave participants points for their performance during training. Lack of an effect of outbreak severity may indicate that the point system was ineffective. Participants may have lacked the motivation to use it: it did not act as a basis for monetary payment. Alternatively, they may have found it difficult to interpret: points received after a given percentage error were inversely related to outbreak severity, and people find inverse relationships such as this difficult to learn (e.g. Brehmer, 1973; Brehmer, Kuylenstierna, & Liljergren, 1974). In fact, during the informal debriefing after the experiment, participants occasionally admitted that they paid little attention to the points that they were given. However, they still claimed that they recognized that more severe outbreaks would be more important within our scenario.

Another possibility needs to be considered. Both of our alternative hypotheses were based on the notion that people initially assess the importance of their task and then make a further decision about how to proceed on the basis of the results of this assessment. Assessing task importance requires judging outbreak severity. But if people can assess outbreak severity, they do not need to make a further decision about how to proceed because assessing outbreak severity is all that their task requires: outbreak severity is given by number of cattle deaths, and number of cattle deaths is the criterion that they are required to estimate. This line of reasoning suggests that lack of an effect of task importance on advice-taking should not be regarded as surprising: people who know they can assess outbreak severity can recognize they do not need advice; those who know they cannot assess it can recognize that they need advice but cannot tailor the amount of advice they take to task importance.

In fact, the circularity of the argument here is only apparent. It disappears once it is recognized that ability to assess outbreak severity need not be all-or-none: it can be partial. Consider, for example, people who are only able to forecast outbreak severity as falling within a broad range. If that range is centered in the higher part of the severity continuum, such people may feel that the judgment is important enough to take advice in order to obtain greater precision. On the other hand, if the range is centered in the lower part of severity continuum, they may feel that the consequences of error are low enough for them to accept responsibility for them themselves.

It remains true, however, that people would have to possess some initial level of forecasting competence for any effect of task importance to appear. If they could not even judge outbreak severity within some broad range, they could not use the location of that range as a basis for their decision about whether to take advice. It is quite possible that we failed to obtain an effect of task importance because the 30 training trials that we gave our participants were insufficient to enable them to estimate a range of outbreak severities that was narrow enough to define their judgment as relatively important or unimportant. Perhaps, if our judges had been semi-experienced rather than novices, task importance would have affected advice-taking. With this possibility in mind, we performed another experiment. This time, participants received 100 training trials: from their point of view, their training was equivalent to that of their semi-experienced advisors.

EXPERIMENT 2

People were given 100 training trials. In all other respects, the experiment was identical to Experiment 1. Research on the effectiveness of persuasive messages suggests that the more highly trained judges in this experiment will take less advice than the novices in the first experiment, and that the amount they do take will be less influenced by the level of training attributed to their advisors (e.g., Petty *et al.*, 1981). Given the arguments outlined in the previous section, we were also interested to discover whether outbreak severity would have an effect on the degree to which judges in this experiment took advice.

Method

Participants. Thirty people from the same population as before acted as participants. None of them had taken part in the first experiment. The experiment took each of them about 75 min to complete.

Stimuli, design, and procedure. Participants were given 100 training trials, and the instructions that they received reflected this. In all other respects, the experimental method was identical to that used in the first experiment.

Results

Data were subjected to the same filter as before to eliminate responses that were clearly the result of mistypings. Again, this resulted in fewer than 1% of the data being excluded from training and test sessions.

Figure 2a shows that MAPE scores decreased over

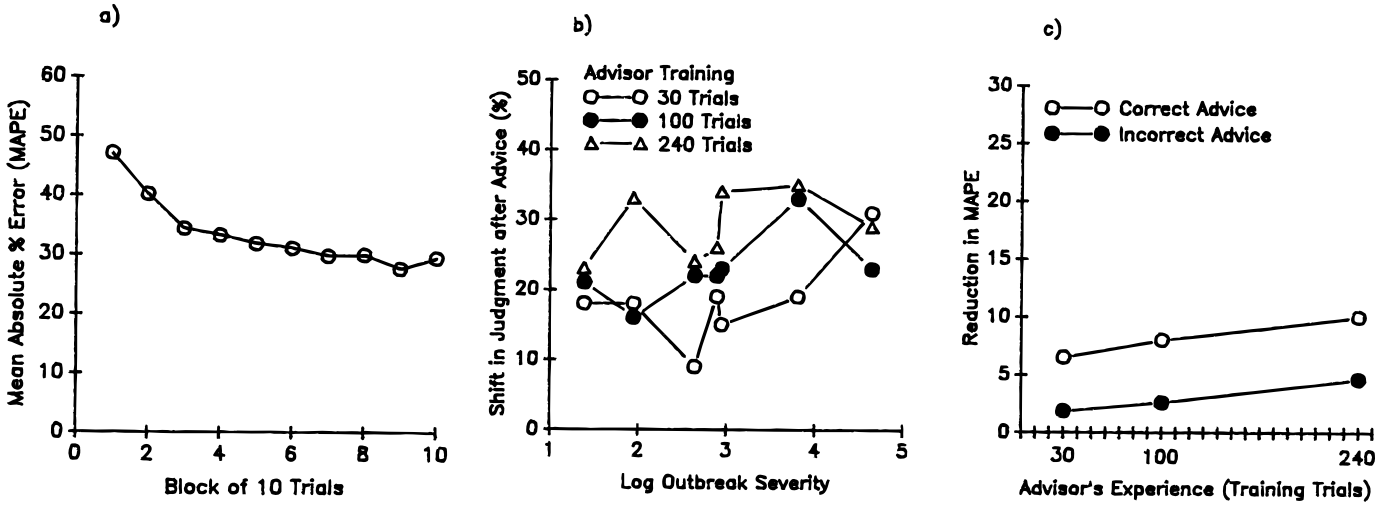

FIG. 2. Experiment 2. (a) Mean absolute percentage error (MAPE) over 10 blocks of 10 trials in the training session. (b) Percentage shift in judgment after advice. Use of log scale on abscissa is to aid visual clarity. (c) Reduction in MAPE after advice.

the 10 blocks of the training session. Slope of the graph of the log of their standard deviations against the log of their means was 0.02, indicating that data transformation was unnecessary. Both linear ($t(1) = 4.98$; $p < .001$) and quadratic ($t(1) = 3.47$; $p < .01$) coefficients of the regression of MAPE scores on to trial blocks were significant, and the overall fit of this model was good ($F(2,9) = 31.47$; $p < .001$; adj. $R^2 = .87$). Thus learning took place and showed evidence of approaching its asymptotic level. Performance on the third block in this experiment was not significantly different from performance on the third block in the previous one. However, performance on the last block of this experiment was significantly better than performance on the last block of the previous one ($t(52) = 3.23$; $p < .01$). Thus the additional training was effective in producing additional learning.

For each trial in the test session, percentage shift between pre-advice and post-advice judgments was calculated as before. Figure 2b shows the mean values of these shifts as a function of level of training attributed to advisors and outbreak severity. Inspection of the figure suggests that shifts again increased with level of training attributed to advisors, and that there may be some effect of outbreak severity. An analysis of variance using advisors' training level, outbreak severity, and advice correctness as within-subject variables confirmed the first of these impressions but not the second one. Effect of advisors' training level was highly significant ($F(2,2062) = 10.90$; $p < .0001$) but neither of the other main effects nor any of the interactions reached significance.

Figure 2c shows reduction in MAPE scores between pre-advice and post-advice judgments as a function of the training attributed to the advisors and the quality of advice they gave. We performed an analysis of variance on these scores using advisors' training level, outbreak severity, advice correctness, and occasion (pre- vs post-advice) as within-subject variables. However, as the graph of the log of the cell standard deviations against the log of their means had a slope of .89, we first subjected them to a logarithmic transformation (Box & Cox, 1964).

The pattern of results produced by this analysis was quite similar to that in the first experiment. The effect of occasion was highly significant ($F(1,4210) = 10.47$; $p < .01$), showing that advice improved judgments. Even poor advice produced some improvement. However, the greater improvement produced by correct advice led to a main effect of advice correctness ($F(1,4210) = 5.58$; $p < .05$) and an interaction between this variable and occasion ($F(1,4210) = 19.82$; $p < .0001$). Given that people took more advice from more highly trained advisors and that advice benefited their judgments, one would expect the reduction in MAPE to be greater when advice was taken from more experienced advisors. Figure 2c appears to show some evidence of such an effect but it failed to reach significance.

MAPE scores were higher and affected more by advice correctness and by the different amounts of advice taken from different types of advisor for the two least severe outbreaks than for the other six outbreaks. These differences produced a main effect of severity ($F(7,4210) = 53.70$; $p < .0001$) and interactions between this factor and advisors' training level ($F(14,4210) = 2.42$, $p < .01$), between it and advice correctness ($F(7,4210) = 2.19$; $p < .05$), and between all three of these variables ($F(14,4210) = 2.56$; $p < .01$). Again, these effects can be interpreted as artifacts of the small absolute size of the criterion values in the two least

severe outbreaks and of the experimental requirement to produce judgments as whole numbers.

Discussion

Superficially, the results of this experiment appear very similar to those of the first one. The mean shift in judgment after advice was lower (23%) than before (30%) but level of advisor training was still the only variable that significantly affected the size of this shift. However, whereas the size of the shift was identical for the least severe and most severe outbreaks in the first experiment (31%), it was 7% higher for the most severe outbreak in the present experiment. We suspect that there was a real but small effect of severity but that our experimental design lacked the statistical power required to reveal it as significant. Rather than repeat the experiment with a larger group of participants, we decided to modify it in a way that we thought would increase the size of the effect to be obtained.

EXPERIMENT 3

Perhaps most people need more than 100 training trials to estimate a range of outbreak severities narrow enough to define the importance of their task and hence to act as a basis for deciding how much advice to take. With this possibility in mind, we designed a third experiment. The amount of training we gave judges was more than doubled to 240 trials to ensure that any embryonic effect of outbreak severity in the second experiment would emerge fully fledged as a significant one in this experiment. This means that, from their point of view, they had received the same amount of training as their experienced advisors and considerably more than either their semi-experienced or their novice advisors.

Based on work on the effectiveness of persuasive messages (e.g., Petty *et al.*, 1981), we also expected this additional training to reduce the shift in judgment caused by the advice and to lessen the influence that the level of training attributed to advisors has on the size of this shift.

Method

Participants. Thirty more people from the same population as before acted as participants. None of them had participated in either of the earlier experiments. They each required about 90 min to complete the training and test sessions.

Stimuli, design, and procedure. Participants received 240 training trials, and the instructions that they were given reflected this. In all other respects, the experimental method was identical to that in the first two experiments.

Results

Data were subjected to the same filter as before to exclude responses that were clearly the result of mistypings. Again, this resulted in no more than 1% of the data being dropped from the training and test sessions.

Figure 3a shows that MAPE scores decreased over the 24 blocks of the training session. Slope of the graph of the log of their standard deviations against the log of their means was 0.08, indicating that data transformation was unnecessary. Both linear ($t(1) = 6.41$; $p < .001$) and quadratic ($t(1) = 5.08$; $p < .001$) coefficients of the regression of MAPE scores on to trial blocks were significant, and the overall fit of this model was good ($F(2,23) = 31.84$; $p < .001$; adj. $R^2 = .73$). Thus learning took place and showed evidence of becoming asymptotic. Performance on the third block in this experiment was not significantly different from performance on that block in either of the previous experiments. Neither was performance on the tenth block of this experiment significantly different from performance on that block in the second experiment. Finally, performance on the last block of the present experiment was significantly better than performance on the last block of the first experiment ($t(51) = 2.08$; $p < .05$) but not of the second experiment. Thus increasing the level of training from 10 to 24 blocks of trials failed to produce a further reduction in forecast error.

As before, percentage shift between pre-advice and post-advice judgments was calculated for each trial of the test session. Figure 3b shows mean values of these shifts as a function of level of training attributed to advisors and outbreak severity. Analysis of variance using advisors' training level, outbreak severity and advice correctness as within-subject variables revealed that only the effect of outbreak severity reached significance ($F(7,2060) = 2.52$; $p < .05$). Whereas mean shifts for the six least severe outbreaks were fairly constant (range 17–19%), that for the most severe outbreak was much larger (35%). Neither of the other main effects nor any of the interactions approached significance.

Figure 3c shows reduction in MAPE scores between pre-advice and post-advice judgments as a function of the training attributed to the advisors and the quality of the advice they gave. We performed an analysis of variance on these scores using advisors' training level, outbreak severity, advice correctness, and occasion (pre- vs post-advice) as within-subject variables. However, as the graph of the log of the cell standard deviations

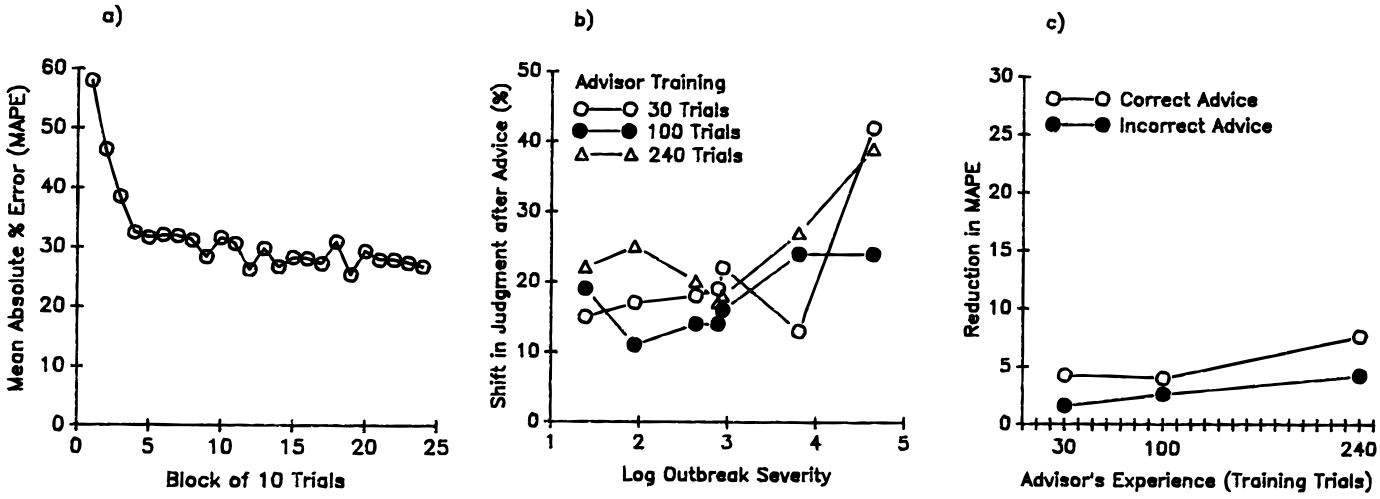

FIG. 3. Experiment 3. (a) Mean absolute percentage error (MAPE) over 24 blocks of 10 trials in the training session. (b) Percentage shift in judgment after advice. Use of log scale on abscissa is to aid visual clarity. (c) Reduction in MAPE after advice.

against the log of their means had a slope of .80, we first applied a logarithmic transformation to them (Box & Cox, 1964).

The effect of occasion was significant ($F(1,4188) = 5.42$; $p < .05$), showing that advice improved judgments. As in previous experiments, an interaction of this variable with advice correctness showed that correct advice was more beneficial ($F(1,4188) = 5.22$; $p < .05$).

There was a significant main effect of severity ($F(7,4188) = 65.76$; $p < .0001$). As before, this effect arose because MAPE scores for the two least severe outbreaks were higher. It can again be interpreted as an artifact of the small absolute size of criterion values in the two least severe outbreaks and of the experimental requirement to give judgments as whole numbers.

Cross-experimental comparison of shifts in judgment after advice. Inspection of Figs. 1b, 2b, and 3b suggests that more highly trained judges were less influenced by advice. This is something to be expected on the basis of research that has been done on the effectiveness of persuasive messages. To discover whether the effect is a significant one, we carried out an analysis of variance with judge's level of training as a between-subjects variable and advisor's level of training and outbreak severity as within-subjects variables.

The analysis showed that judge's level of training did indeed affect the extent to which judgments were influenced by advice ($F(2,87) = 3.92$; $p < .05$). However, as expected on the basis of the individual analyses of the three experiments, a significant main effect of advisor training ($F(2,174) = 23.56$; $p < .0001$) and a significant interaction between judge's training and advisor's training ($F(4,174) = 5.35$; $p < .001$) were also obtained. As Fig. 4a shows, all judges, whatever their level of training, shift their judgments about 20% when presented with advice from someone who has had little training. Highly trained judges still shift their judgments by this amount when the advice comes from someone who is moderately or highly trained. However, judges who are less well trained shift their judgments more when their advisors are better trained than they are. This increase is greater when the difference between judges' and advisors' training levels is higher.

The analysis also produced a significant effect of outbreak severity ($F(7,609) = 2.57$; $p < .05$). Figure 4b shows a shift in judgment after advice as a function of judge's training and outbreak severity. Although the interaction between these two variables failed to reach significance, we have seen from our individual analyses of the three experiments that the change in shift as outbreaks become more severe was negligible when judges were poorly trained, positive but nonsignificant when they were moderately trained, and significant when they were highly trained.

Discussion

In this third experiment, we increased the level of training that judges received to the level attributed to their most experienced advisors. Although this did not improve their performance, it is likely to have increased their confidence (Harvey, 1994). This, in turn, would have influenced their advice-taking (Arkes *et al.*, 1987; Trafimow & Sniezek, 1994). More specifically, work on persuasion (Wood, 1982; Wood *et al.*, 1985) suggests that they would become less influenced by information about the credibility of the source of their advice. This is exactly what we found: the effect of level of training attributed to advisors was eliminated.

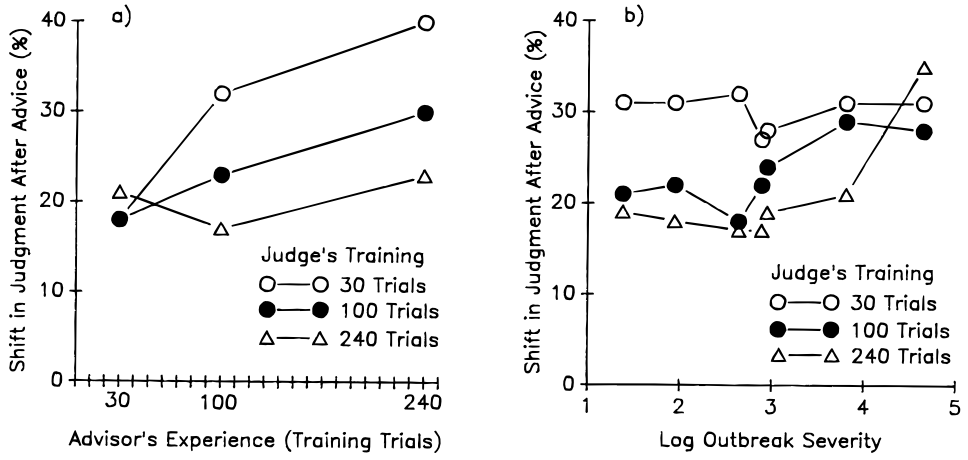

FIG. 4. (a) Percentage shift in judgment after advice as a function of judge's training and advisor's training. (b) Percentage shift in judgment after advice as a function of outbreak severity and judge's training.

Although the extra experience that people received in this experiment did not significantly reduce their overall forecast error, it did enable them to appreciate and be influenced by the different degrees of task importance.[4] But why were they influenced by task importance? Earlier, we considered two possibilities. First, people may be able to improve their own judgments by putting more effort into them. If they could, they would be more likely to do so when they perceived the task as important (Payne *et al.*, 1993). This would have resulted in less advice-taking when outbreaks were severe. This is opposite to the pattern of results that we have obtained here.

The second possibility was that people recognize advice as potentially beneficial but find integrating it with their own judgments cognitively demanding. Hence they would be more likely to make the effort to use the advice when the task is more important and when the advice comes from a more credible source. We did indeed find that advice-taking was greater when the task was more important. However, as we have seen, there was no evidence that it was greater when its source was more credible.

If people were taking advice for no other reason than to improve their judgments, they should have weighted advice from experienced sources more heavily. That this pattern did not emerge here suggests that the need to improve judgment was not the only reason that people took advice. In what follows, we propose that another quite different factor was responsible for the effect of task importance obtained in this third experiment.

[4] That people in this experiment continued to learn something new about the task during their additional training trials and used what they had learnt during their test trials implies that the above-zero asymptote seen in Fig. 3a is not symptomatic of mental fatigue. It is, instead, likely to be related to the fact that cues were presented pictorially rather than numerically (e.g. Cooksey, 1996).

GENERAL DISCUSSION

The three experiments produced a fairly complex pattern of results. However, we shall suggest here that the amount of advice people take can be explained fairly simply in terms of three separate components. First, it appears that there is a basic level of advice-taking that is present whatever the other features of the task. Second, when judges believe advisors have greater expertise than they do themselves, they add a component to this basic level. The size of the component they add is related to the extent to which the advisor appears to have greater expertise. The final component is involved only when judges acquire sufficient experience to distinguish important judgments from relatively unimportant ones.

In the following sections, we discuss each of these three components in turn. We tentatively identify them with accepting help, improving judgment, and sharing responsibility for high-risk judgments.

Accepting Help

Figure 4a shows that people shifted their judgments about 20% towards advice that they believed had come from someone who had received no more training than they had themselves. Even experienced judges (240 training trials) shifted their judgments by this amount in response to advice from novices (30 training trials) who, they were told, had received eight times less training than they had themselves. Why did these experienced participants take any account of the novices' views?

One possibility is that they recognized that combining forecasts of different quality can reduce error variance and hence produce a prediction that is more accurate than either forecast separately. However, the

possible advantages of combination for the better forecaster are still less when the other forecaster is worse. But, as we have seen, experienced judges took as much account of advice from novices as they did of that from people as highly trained as themselves.

A second possibility is that experienced judges (but not novices or semi-experienced ones) came to view the improvement due to training as having reached asymptote by the time that 30 training trials had been given. This could explain why they shifted their judgments in response to advice, but did so by the same amount whatever the training level of their advisors. However, according to this view, they should have taken the average of their initial judgment and the advice, thus producing a shift of 50% rather than 20%.

A third possibility is that experienced judges (but not others) were able to recognize that the advice they were told came from advisors with different degrees of training was, in fact, of the same quality. However, analysis of the third experiment failed to show that experienced judges were any more capable of distinguishing correct advice from incorrect advice than the less well trained judges in the first two experiments. If they were not sensitive to advice correctness, it is difficult to see how they could recognize what other judges could not (viz. advice was no better when it came from advisors attributed with higher levels of training).

A final possibility is that the very act of offering someone advice puts that person under social pressure to comply with it. Advice is usually given in the spirit of helping someone to do something or of helping them learn to do it. People may be reluctant to ignore advice because they are concerned that this will be taken as a rejection of help freely offered.[5] Even if they believe the advice to be without value, they may shift their judgment a token amount toward it. The 20% shift that we observed judges to make when they believed their advisors to have received no more training than they had themselves may represent this token amount.

Sniezek and Buckley's (1995) experiment convincingly demonstrates the importance of this type of social compliance for advice-taking. Their judges answered two-choice questions after being provided with advice from two other people. Both judges and advisors provided estimates of the probabilities that their choices were correct by choosing numbers on a scale that ranged from .50 to 1.00. On 54 occasions, the two advisors selected the same answer to the question but, via .50 confidence estimates, signaled to the judge that they both regarded their choice as a pure guess and hence, presumably, valueless as advice. Despite this, judges accepted their advice on 47 (87%) of the 54 occasions. The 20% shift that we observed when advisors were no better trained than judges is not so striking as this, but then our judges were not in direct social contact with their advisors.

The amount of advice-taking that is related to concerns about the social unacceptability of complete rejection of advice may depend on contextual factors. It was constant at about 20% within our experimental paradigm. In other situations, such as where there is direct social contact between judges and advisors, it may well be higher. Conversely, when collecting advice incurs a monetary cost, it is likely to be lower (cf. Connolly & Wholey, 1988). Finally, we would also expect it to vary considerably from individual to individual in a manner that may depend on personality factors. Unfortunately, whereas there is a large body of work on factors that determine how much help people offer (e.g., Latané & Nida, 1981), there is a paucity of research on factors that determine how much help people accept.

Improving Judgment

Figure 4a shows that people who believed that their advisor had been provided with more training than they had received themselves shifted their judgments towards the advice by more than the basic 20% that we have associated with an unwillingness to reject help. Furthermore, they related the size of this additional advice-taking to the apparent advantage in judgment skill that their advisor had over them. This broad pattern of results is what would be expected on the basis of Hogarth and Einhorn's (1992) Belief Adjustment model. It is also consistent with previous work on information combination (e.g., Birnbaum *et al.*, 1976) and persuasion (e.g. Petty & Cacciopo, 1986).

These features of the results can be interpreted in terms of participants' efforts to improve the quality of their judgments. Not unreasonably, they assume that advisors who are more highly trained than they are themselves will provide them with real help and that this help will be greater when their advisors have much more training than when they have only a modest amount more.

There is, however, one feature of the results that initially appears anomalous if people were using better-trained advisors to provide them with real help in improving the quality of their judgments. Even when they believed their advisors had been given eight times more training than they had received themselves, the mean shift in their judgment towards the advice was still only 40%. They did not even take a simple average of their own judgment and the advice: despite the difference in

[5] More generally, people may be reluctant to reject help that is not needed because this may reduce the chances of it being offered in the future when it is needed.

training, they weighted the former more heavily. Why was this?

One possibility is that people used their own initial judgment as an anchor and that they adjusted away from it insufficiently when presented with advice. Such under-adjustment is typically observed when people employ anchor-and-adjust heuristics in judgment tasks (e.g. Tversky & Kahneman, 1974). Although it is difficult to discount this explanation, there are problems with it. For example, we saw in the last section that experienced judges take as much account of the views of novices as they do of those of their peers: they appear to over-adjust rather than under-adjust when presented with advice from novices. Why did they not use the anchor-and-adjust heuristic or, if they did, why were social-compliance effects more than sufficient to counteract underadjustment in judgments that they made but not in judgments that novices made?

Novices had no experience of being highly trained: they had no real basis for assessing the value of experienced judges' advice. They were free to doubt the value of any training additional to the little that they had received. In other words, they were provided with scope for being optimistic about the relative quality of their judgments compared with those made by people with more experience. Perhaps this accounts for why they were not influenced more by advice from experienced judges.

Superficially, this explanation seems wanting. Even if novices viewed *any* training over and above the 30 trials that they had received as worthless, they should still have valued judgments of experienced people as equal to theirs; in fact, as we have seen, they valued their own more. Furthermore, it is clear from the results of Experiment 1 (Fig. 1b) that novices did not consider the continuation of training beyond what they had received as worthless; they took more account of advice from people who were more highly trained.

It is possible, however, that novices recognized the value of training *per se* but used their scope for being optimistic to overestimate the value of training to them personally relative to its value for other people. This possibility is suggested by a number of studies. Svenson (1981) and McKenna (1993) have shown that, on average, car drivers judge their skill levels as higher and their chances of being in an accident as lower than those of the average driver. More recently, Koehler and Harvey (1997) found that actors were more overconfident in their performance in a judgmental control task than people who merely observed that same performance. Clearly, then, participants in the present experiments may have recognized that providing advisors with more than 30 training trials would improve their advice but may still have reckoned that their own judgments would be better than those of advisors who had received more training than they had.

Confirmation of judges' overconfidence in their own judgments relative to those of advisors is reported by Gardner and Berry (1995). People had to bring the outputs of a dynamical system into target ranges by altering the value of control parameters. Advice was available to them. The advice was always correct: "It was made clear to . . . experimental groups that no tricks were involved and that no wrong advice would be given" (p. 563). Furthermore, advice was free: taking it incurred no monetary penalty. Unsurprisingly, taking advice improved performance. Yet despite the fact that advice was beneficial to judgment, free and known to be correct, people were reluctant to accept it fully. For example, in one experiment, advice was presented automatically and specified whether each control parameter should be moved up, moved down or maintained at its current value: only 73% of participants acted in full agreement with the advice. In another experiment, advice was presented only when people asked for it: they requested it for only 44% of their judgments.

Sharing Responsibility

The effect of judgment importance on advice-taking is different from the effect of issue importance in persuasion. First, it results in people being more rather than less influenced by the views of others. Second, it is determined not by the levels of expertise that they perceive those other people to have but by the levels of expertise that they perceive themselves to have.

When judges were novices (Experiment 1), their advice-taking appears to have been determined only by their desire not to reject help and by their desire to improve the accuracy of their forecasts. As Fig. 4b shows, the percentage shift in their judgment after advice was fairly constant at about 30% across all severity levels: this value is, of course, the average of the three points in the upper curve of Fig. 4a.

When judges were experienced (Experiment 3), their advice-taking for all but the most important judgment appears to have been determined solely by their desire not to reject all help. For all these judgments, advice-taking was at the basic level attributed to that component. However, for the most important judgment, it was about twice as high (Fig. 4b). It appears that another component, not related to the perceived expertise of the advisor, came into play for this judgment.

Semi-experienced judges (Experiment 2) were midway between novices and experienced judges. Like the other two groups, their advice-taking was partly determined by their desire not to reject all help. Like novices,

it was also determined by their opportunity to improve their judgments by taking more advice from advisors who had received more training than they had themselves (Fig. 4a). Finally, there was some suggestion that, like experienced judges, they took more advice for the more important judgments (Fig. 4b). However, this last effect differs from the corresponding one found for experienced judges in two ways. First, it was smaller and more variable: it did not reach significance. Second, advice-taking appears to be relatively high not just for the most important judgment but for the two or three most important ones (Fig. 4b). What were the reasons for these differences?

People had to be able to assess importance before an importance-dependent effect could appear in the data. Judgment importance was determined by outbreak severity. As expertise in the task was defined by ability to assess outbreak severity, it is not surprising that the effect of importance evident when people were experienced was only half-evident when they were semiexperienced.

Why would the minimum level of judgment importance for which advice-taking was elevated be lower for semiexperienced judges than for experienced ones? One possibility is that taking advice allows people to share responsibility for the consequences of error but that they decide to share responsibility only when the risk associated with the judgment exceeds some threshold level.

Risk is often characterized as expected loss (e.g., Yates & Stone, 1993). In other words, it can be regarded as a product of the negative consequences arising from an undesirable outcome and the probability of that outcome. In our experiments, the undesirable outcomes were errors in judgment. Consider one such outcome: an error of 20% in the forecast. People would perceive negative consequences of such an error to increase with outbreak severity and the probability of one to decrease with their experience at the task. Thus, if they need to share responsibility for judgment when risk exceeds some threshold level, the minimum level of outbreak severity for which they do try to share it should be lower when they are less well trained (Fig. 4b).

According to this account, people add a component to their advice-taking that represents a sharing of responsibility for the judgment when the risk associated with error is high. Their ability to add this component in the present task depended on their experience in the task; the level of judgment importance at which they added it depended on their assessment of their own expertise. The lack of an interaction between the effects of importance and advisor training is not surprising because, for the purposes of sharing responsibility, anyone will do, regardless of their level of experience.

This account of the importance effect is plausible but speculative. To test it, it would be useful to employ monetary pay-offs and to combine them with a means of making the responsibility of advisors explicit. For example, the importance of judgments could be defined in terms of the size of the gains received when they are correct and the size of the losses suffered when they are in error. Proportion of gains and losses assigned to advisors could then be specified as a proportion of the percentage shift in judgment that follows their advice. If people attempt to share responsibility for their judgments when risk reaches some threshold level, increasing this proportion should increase the minimum level of judgment importance for which their advice-taking incorporates the responsibility-sharing factor.

Finally, it is worth pointing out that the notion that advice-taking allows people to share responsibility for important judgments and decisions chimes with the views of those policy researchers who have argued that governments use commissions of inquiry not only to identify appropriate courses of action but also to legitimize action or inaction (e.g., Bulmer, 1993).

Summary: Taking Advice

We have argued that people considering how much to be influenced by advice have three aims in mind: they want to avoid completely rejecting help that is offered to them; they want to improve the quality of their judgments; they want to share responsibility for high-risk judgments.

People's reluctance to reject help is indicated by their willingness to take some account of the views of people whom they believe to be very much less expert at the task than they are. Their insensitivity to differences in the levels of training of these less well trained people suggests that they are not taking their views into account merely to reduce error variance in their judgments.

People improve their judgments by taking advice from those who they perceive to be more expert than they are themselves and by taking more advice from those who they perceive to have greater expertise. However, their ability to use advice to improve judgments appears constrained by their overestimation of their own judgment skill relative to that of people who have had as much or more training.

As people gain experience in the task, they become able to distinguish high-risk judgments from others. They share responsibility for these judgments by taking about twice as much advice for them as they do for the others. There is some suggestion that less well trained

judges attempt to share responsibility for less important judgments: this would be expected if the threshold of risk that triggers responsibility-sharing is independent of the experience that the judge has in the task.

This three-component model of advice-taking is presented as an attempt to explain the pattern of results that we obtained in a situation in which advice was freely available to people. It would be interesting to examine how far its applicability extends to conditions in which advice-taking is a two-stage process: advice collection (costing time, money, or effort) and advice use. The three components that we have identified may influence both of these stages or affect just one of them.

How does our model relate to Hogarth and Einhorn's (1992) Belief Adjustment model? The second of our three components (improving judgment) could operate in a way similar to that which they suggest. However, we propose that additional components (accepting help, sharing responsibility) come into play when there is a social dimension to the adjustment process.

REFERENCES

Adams, J. A., & Goetz, E. T. (1973). Feedback and practice as variables in error detection and correction. *Journal of Motor Behavior*, **5,** 217–224.

Adelman, L. (1981). The influence of formal, substantive, and contextual task properties on the relative effectiveness of different forms of feedback in multiple-cue probability learning tasks. *Organizational Behavior and Human Performance*, **27,** 423–442.

Ahlburg, D. A. (1992). Error measures and the choice of a forecast method. *International Journal of Forecasting*, **8,** 99–100.

Arkes, H. R., Christensen, C., Lai, C., & Blumer, C. (1987). Two methods of reducing overconfidence. *Organizational Behavior and Human Decision Processes*, **39,** 133–144.

Armstrong, J. S., & Collopy, F. (1992). Error measures for generalizing about forecasting methods: Empirical comparisons. *International Journal of Forecasting*, **8,** 69–80.

Berry, D. C., & Broadbent, D. (1984). On the relationship between task performance and associated verbalizable knowledge. *Quarterly Journal of Experimental Psychology*, **36A,** 209–231.

Birnbaum, M. H., & Stegner, S. E. (1979). Source credibility in social judgment: Bias, expertise and the judge's point of view. *Journal of Personality and Social Psychology*, **37,** 48–74.

Birnbaum, M. H., Wong, R., & Wong, L. K. (1976). Combining information from sources that vary in credibility. *Memory and Cognition*, **4,** 330–336.

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B*, **26,** 211–252.

Brehmer, B. (1973). Single-cue probability learning as a function of the sign and magnitude of the correlation between cue and criterion. *Organizational Behavior and Human Performance*, **9,** 377–395.

Brehmer, B., Kuylenstierna, J., & Liljergren, J.-E. (1974). Effects of function form and cue validity on the subjects' hypotheses in probabilistic inference tasks. *Organizational Behavior and Human Performance*, **11,** 338–354.

Bulmer, M. (1993). The Royal Commission and departmental committee in the British policy-making process. In B. G. Peters & A. Barker (Eds.), *Advising West European governments: Inquires, expertise and public policy*. Edinburgh: Edinburgh University Press.

Carbone, R., & Armstrong, J. S. (1982). Evaluation of extrapolative forecasting methods: Results of a survey of academicians and practitioners. *Journal of Forecasting*, **1,** 215–217.

Chaiken, S., & Maheswaran, D. (1994). Heuristic processing can bias systematic processing: Effects of source credibility, argument ambiguity and task importance on attitude judgment. *Journal of Personality and Social Psychology*, **66,** 460–473.

Chatfield, C. (1988). Editorial. Apples, oranges and mean square error. *International Journal of Forecasting*, **4,** 515–518.

Connolly, T., & Wholey, D. R. (1988). Information mispurchase in judgment tasks: A task-driven causal mechanism. *Organizational Behavior and Human Decision Processes*, **42,** 75–87.

Edgell, S. E. (1993). Using configural and dimensional information. In N. J. Castellan, Jr. (Ed.), *Individual and group decision making* (pp. 43–64). Hillsdale, NJ: Erlbaum.

Fildes, R. (1992). The evaluation of extrapolative forecasting methods. *International Journal of Forecasting*, **8,** 81–98.

Gardner, D. H., & Berry, D. C. (1995). The effect of different forms of advice on the control of a simulated complex system. *Applied Cognitive Psychology*, **9,** 555–579.

Hammond, K. R., & Summers, D. A. (1965). Cognitive dependence on linear and nonlinear cues. *Psychological Review*, **72,** 215–224.

Harvey, N. (1994). Relations between confidence and skilled performance. In G. Wright and P. Ayton (Eds.), *Subjective probability* (pp. 321–352). New York: Wiley.

Harvey, N., & Bolger, F. (1996). Graphs versus tables: Effects of data presentation format on judgmental forecasting. *International Journal of Forecasting*, **12,** 119–137.

Hell, W., Gigerenzer, G., Gauggel, S., Mall, M., & Müller, M. (1988). Hindsight bias: An interaction of automatic and motivational factors? *Memory and Cognition*, **16,** 533–538.

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The Belief Adjustment Model. *Cognitive Psychology*, **24,** 1–55.

Koehler, D., & Harvey, N. (1997). Confidence judgments by actors and observers. *Journal of Behavioral Decision Making*, in press.

Latané, B., & Nida, S. (1981). Ten years of research on group size and helping. *Psychological Bulletin*, **89,** 307–324.

Lee, J.-W., & Yates, J. F. (1992). How quantity judgment changes as the number of cues increases: An analytical framework and review. *Psychological Bulletin*, **112,** 363–377.

McKenna, F. P. (1993). It won't happen to me: Unrealistic optimism or illusion of control? *British Journal of Psychology*, **84,** 39–50.

Makridakis, S., Wheelwright, S. C., & McGee, V. (1983). *Forecasting: methods and applications* (2nd ed.). New York: Wiley.

Marteau, T. M., Johnston, M., Wynne, G., & Evans, T. R. (1989). Cognitive factors in the explanation of the mismatch between confidence and competence in performing basic life support. *Psychology and Health*, **3,** 172–182.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker*. Cambridge: Cambridge University Press.

Petty, R. E., & Cacciopo, J. T. (1986). The elaboration likelihood model of persuasion. In *Advances in Experimental Social Psychology* (Vol. 19, pp. 123–205). New York: Academic Press.

Petty, R. E., Cacciopo, J. T., & Goldman, R. (1981). Personal involvement as a determinant of argument-based persuasion. *Journal of Personality and Social Psychology*, **41,** 847–855.

Sanderson, P. (1989). Verbalizable knowledge and skilled task performance: Associations, dissociations, and mental models. *Journal of*

Experimental Psychology: Learning, Memory and Cognition, **15,** 729–747.

Schmitt, N., Coyle, B. W., & Saari, B. B. (1977). Types of task information in multiple-cue probability learning. *Organizational Behavior and Human Performance,* **18,** 316–328.

Sniezek, J. A., & Buckley, T. (1993). Becoming more or less uncertain. In N. J. Castellan, Jr. (Ed.), *Individual and group decision making.* Hillsdale, NJ: Erlbaum.

Sniezek, J. A., & Buckley, T. (1995). Cueing and cognitive conflict in judge-advisor decision making. *Organizational Behavior and Human Decision Processes,* **62,** 159–174.

Stanley, W. B., Mathews, R. C., Buss, R. R., & Kotler-Cope, S. (1989). Insight without awareness: On the interaction of verbalization, instruction and practice in a simulated process control task. *Quarterly Journal of Experimental Psychology,* **41A,** 553–577.

Stewart, T. R. (1988). Judgment analysis: Procedures. In B. Brehmer & C. R. B. Joyce (Eds.), *Human judgment: The SJT view* (pp. 43–64). Amsterdam: Elsevier North-Holland.

Stewart, T. R., & Lusk, C. M. (1994). Seven components of judgmental forecasting skill: Implications for research and the improvement of forecasts. *Journal of Forecasting,* **13,** 579–599.

Received: May 7, 1996

Summers, D. A., & Hammond, K. R. (1966). Inference behavior in multiple-cue tasks involving both linear and nonlinear relations. *Journal of Experimental Psychology,* **71,** 751–757.

Svenson, O. (1981). Are we all less risky and more skilful than our fellow drivers are? *Acta Psychologica,* **47,** 143–148.

Trafimow, D., & Sniezek, J. A. (1994). Perceived expertise and its effects on confidence. *Organizational Behavior and Human Decision Processes,* **57,** 290–302.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science,* **185,** 1124–1131.

Wheelwright, S. C., & Makridakis, S. (1985). *Forecasting methods for management* (4th ed.). New York: Wiley.

Wood, W. (1982). Retrieval of attitude relevant information from memory: Effects on susceptibility to persuasion and on intrinsic motivation. *Journal of Personality and Social Psychology,* **42,** 798–810.

Wood, W., Kallgren, C., & Priesler, R. (1985). Access to attitude relevant information in memory as a determinant of persuasion. *Journal of Experimental Social Psychology,* **21,** 73–85.

Yates, J. F., & Stone, E. R. (1992). The risk construct. In J. F. Yates (Ed.), *Risk-taking behavior.* New York: Wiley.