ORGANIZATIONAL BEHAVIOR AND HUMAN DECISION PROCESSES
Vol. 70, No. 3, June, pp. 267–282, 1997
ARTICLE NO. OB972710

Judgments of Decision Effectiveness: Actor–Observer Differences in Overconfidence

Nigel Harvey

University College London, United Kingdom

Derek J. Koehler

University of Waterloo, Ontario, Canada

and

Peter Ayton

City University, London, United Kingdom

Subjects playing the role of psychiatrists (actors) engaged in a simulated medical decision-making task in which they attempted to bring the value of a patient indicator variable into a desired range. For each treatment recommended by the actor, both the actor and an observer subject playing the role of a nurse assessed the probability that the treatment would be effective. Both actors and observers were overconfident. Actors were more confident in their treatment recommendations than were observers, but this difference was eliminated when observers were given the opportunity to offer their own alternative recommendation. Under the latter circumstances, actors and observers were equally confident in the actors' decisions but observers were more confident than actors in the observers' decisions. These findings suggest that while control over the outcome of the decision has little influence on actor–observer differences in confidence, feedback regarding this outcome plays a crucial role.

People often have to make a series of decisions in order to bring the output of a system into a target range. Doctors modify drug dosage to bring disease states under control with minimal side effects. Stocktakers alter the size of regular orders to ensure that demand is met without excessive storage costs being incurred. Industrial process controllers regulate features of some ongoing reaction to ensure that an appropriate amount of the desired material is produced within safety constraints. Recently, the processes underlying judgmental control in tasks such as these have been studied by examining performance with simulated systems (e.g., Harvey, 1990a; Kleinmuntz & Thomas, 1987; Moray, Lootstein & Pajak, 1986; Sterman, 1989).

In situations like the ones just outlined, the probability that a proposed control decision will be effective often has to be estimated. This is because the executive decision of whether to implement it must take into account both the expense of changing control parameters and the possibility that the costs of moving further away from the target range may be higher than the benefits of moving closer to it. Who should provide these probability estimates? Should it be an individual involved in formulating and implementing the decision, or should it be someone further removed from the decision-making process?

There are reasons to suppose that independent observers would be better able to estimate the probability of decision success or effectiveness than those involved in formulating and implementing the decisions. Langer (1975), for example, observed that people often suffer from an "illusion of control" over objectively uncontrollable events. This tendency to overestimate degree of control was stronger in subjects' assessments of their own actions than it was the actions of an experimenter

Preparation of this article was supported by a grant from the Economic and Social Research Council of the United Kingdom (R000221383) to the first author and by a grant from the Natural Sciences and Engineering Research Council of Canada (OGP 0183792) to the second author. Parts of this paper were presented at the 31st Annual Meeting of the Psychonomics Society, New Orleans, 1990 and at the 14th Conference on Subjective Probability, Utility and Decision Making, Aix-en-Provence, 1993.

Address correspondence and reprint requests to Derek J. Koehler, Department of Psychology, University of Waterloo, Waterloo, Ontario, N2L 3G1, Canada. E-mail: dkoehler@watarts.uwaterloo.ca.

0749-5978/97 $25.00

(Langer, 1975, Experiment 4). While judgmental control tasks differ from Langer's in that the outcome is objectively under the subject's control, recent research indicates that subjects in such tasks also tend to be overconfident in their ability to control the outcome (Harvey, 1990b).

Langer's (1975, Experiment 4) results suggest that this overconfidence may be lower when the responses that people assess are those of others than when they are their own. Tyebjee's (1987) simulation of new product planning lends to support to this view. He found that subjects who set the level of a marketing budget themselves predicted a higher probability of meeting sales objectives than subjects who had the marketing budget set for them. Work by Wright and Ayton (1989) also suggests that independent observers may be in the best position to forecast decision effectiveness. They found that overconfidence in personal event forecasting correlated with degree of perceived control over the event. For example, overconfidence in 4-week forecasts was greater for personal events, such as losing a checkbook, than for events of an impersonal nature, such as a volcanic eruption.

In many ways, however, individuals who formulate and implement decisions (who we will call *actors*) would seem to be in a better position to judge the probability that those decisions will be effective or successful. Unlike observers, they are forced to attend to the control task in order to perform it and they may, therefore, develop a better understanding of the way the system is responding to their interventions. Also, they know their reasons for formulating particular decisions. Observers deprived of such information may be less able to estimate the probability of efficacy of those decisions.

Some related research, though not directly concerned with the issue of decision control and implementation, provides indirect support for this view. Vallone, Griffin, Lin, and Ross (1990) compared the confidence and accuracy with which undergraduates made predictions about events that might occur during their first year at university, both in their own lives and in their roommates' lives. These researchers found that, while the students were overconfident in both types of predictions, they were no more overconfident in their predictions regarding events in their own lives (over which they presumably had some control) than they were in predictions regarding their roommates. If anything, overconfidence was greater for the latter. Furthermore, Koehler (1994) found that actors were less overconfident than observers in answers to open-ended general knowledge and prediction items, a result that was attributed to the actors' tendency to consider a greater number of alternative possibilities. In summary, then, the evidence from previous research is mixed in its implications regarding actor–observer differences in judgments of decision effectiveness.

To investigate this issue, we conducted a series of experiments in which each actor attempting to control a system was accompanied by an observer. These individuals received the same information about system behavior as the actors and also saw what control decisions were made. After the decision was made, actors and observers independently estimated the probability that it would be successful. Probability estimates made by each member of a pair were not seen by the other member.

The output of the system that had to be controlled in these experiments was determined by a noisy logistic map:

$$Y_{t+1} = AY_t(1 - Y_t) + e_{t+1} \quad (1)$$

where Y is a variable with values between 0 and 1 produced at times $t = 0, 1, 2 \ldots n$; A is a parameter controlled by actors to produce different values of Y; and e is an independently normally distributed error term with a mean of 0 and a variance of .01.

Different types of behavior are produced by the logistic map when the control parameter A is set at different values (Crutchfield, Farmer, & Huberman, 1982; May, 1986). For the noisy system, these different behaviors may be summarized as follows. For A greater than 1.0 but less than 3.0, Y has a single asymptotically stable value that increases with A. For A greater than 3.0 but less than 3.57, the system asymptotes to a stable state in which Y alternates between two fixed values. The difference between these two fixed values increases with A. After A reaches 3.57, the system produces unpredictable chaotic behavior.

In our experiments, actors were initially presented with the system in a stable state in which the behavior of Y was constant, alternating or chaotic. They had to make adjustments to A to ensure that Y was produced with a new constant value. By manipulating the initial behavior of the system (i.e., the starting value of A), we are able to vary task difficulty (Harvey, 1990a) and thereby examine actor-observer differences over a large range of probabilities. Because performance does not generally improve over the session when the initial system parameter value varies, improvements in probability estimation over the course of the experiment cannot be attributed to changes in the probabilities being estimated. This task, then, is an ideal one with which to investigate the influence of factors such as perceived control and outcome feedback on actor–observer differences in assessments of confidence in decision effectiveness.

EXPERIMENT 1

Method

Subjects. Subjects were 60 undergraduate volunteers aged between 18 and 30 years. Each served for one session of approximately 70 min duration.

Procedure. The task, which was run on a computer, was framed as one of medical decision making. Subjects were divided into 30 doctor–nurse pairs. *Doctors* (actors) were told to imagine that they were psychiatrists specializing in drug treatment of affective disorders. *Nurses* (observers) were asked to think of themselves as nurses with an interest in the efficacy of their patients' treatments. Each pair was presented with the same eight patients but in different random orders. For each patient, doctors prescribed 10 treatments and, after each of these, both subjects independently assessed the probability that it would be effective.

Output of the logistic map was multiplied by 50 and rounded to the nearest whole number to produce a diagnostic indicator termed the "happiness index." Subjects were informed that normal people have a monthly index in the range 29–31, and that the aim of treatment was to bring moods into this range. Each of the patients was characterized by a personal value of the control parameter, A. There were two depressives (A = 2.0 and 2.2), one manic (A = 2.9), three alternating manic-depressives (A = 3.1, 3.3 and 3.5), and two chaotic manic-depressives (A = 3.7 and 3.9). Generally speaking, the difficulty of successfully treating the patient increases with the value of A. Subjects treated the patients by prescribing either lithium (which reduced A) or antidepressant (which increased A). Dosage of each drug could vary between 0 and 30 units. Each extra unit moved A a further .05 in the direction specified by the drug. Table 1 shows examples of mood scores for each patient over a period of 12 months when no treatment is given.

Subject pairs were run separately. Members of each pair sat side by side in front of the computer monitor, with doctors on the right and nurses on the left. The top two-thirds of the screen contained information that was to be available to both subjects. The bottom third of the screen was reserved for information confidential to either the doctor or the nurse. To ensure this confidentiality, a T-shaped screen was placed between the subjects perpendicular to the computer monitor in such a way that it occluded the bottom third of the screen on each side from the subject sitting on the other side. On the bench in front of each subject was a joystick console used to enter responses into the computer.

Instructions included the same information for both subjects but emphasized the different roles that each one was to adopt. Thus doctors were told to imagine that they were psychiatrists treating people with mood disorders whereas nurses were asked to imagine that they were assisting the psychiatrists. Both subjects were told:

> In the absence of treatment, some people are consistently more depressed than normal people. They must be treated with an antidepressant drug. The more depressed they are, the more antidepressant must be prescribed to bring their moods into the normal range. Other patients with mood disorders are either manics or manic-depressives. Manics are consistently more irresponsible and boisterous than normal people. Manic-depressives' moods change between excessively boisterous and excessively depressed. In the absence of treatment, some manic-depressives' moods oscillate between the two extremes in a predictable manner, whereas others have quite unpredictable changes in mood. Both manic and manic-depressive patients must be treated with the drug lithium. In treating patients, a psychiatrist must try to find out the most appropriate dosage for bringing each patient's moods into the normal range.

Subjects were told that for diagnostic purposes each patient was asked to keep a diary of experiences that made them feel noticeably happy. The total number of such experiences for a given month was referred to as the "happiness index" for that month. Normal people, subjects were informed, usually have a monthly happiness index of between 29 and 31.

TABLE 1

Examples of Simulated Mood Scores for the Eight Patients in the Absence of Treatment

	Month											
Parameter	1	2	3	4	5	6	7	8	9	10	11	12
2.0	24	25	24	24	23	24	24	25	24	25	24	25
2.2	26	27	26	27	26	27	26	26	26	26	25	26
2.9	32	33	32	32	33	32	32	31	32	32	33	33
3.1	25	38	28	37	27	38	27	38	29	36	28	37
3.3	24	41	25	40	24	41	22	42	23	40	23	40
3.5	18	41	22	42	18	40	22	43	18	40	21	42
3.7	45	15	40	16	38	32	16	41	15	45	26	41
3.9	9	31	45	15	42	24	48	3	12	36	44	18

Subjects were then told that for each patient they would first be presented with happiness indices for two successive months (in the absence of any treatment). The doctor would then make a prescription. After that, doctor and nurse would independently assess the probability that this treatment would bring the patient's happiness index within the normal range (i.e., 29–31) on the following month. Subjects were told that this judgment was important because relatives and other people caring for patients find it useful to have an idea of the extent to which they can expect normal behavior from the patient after each prescription. The system would then step forward to display the patient's happiness index at the end of the third month (i.e. after the first treatment). On the basis of this and knowledge of the previous happiness indices, the doctor was to make a treatment decision for the next month, which both doctor and nurse would once again assess. This cycle of index examination, treatment decision, and assessment of treatment effectiveness would continue until the end of the 12th month. At this point, a new patient would appear for treatment. Subjects were informed that they would see a total of eight patients.

It was explained that if treatment was stopped at any time, the patient would revert to the original pattern of moods that was evident in the first two months before any drugs had been prescribed. They were also informed that for each patient there was a drug dosage that, if maintained, would ensure that the patient's moods were generally within the normal range, though this dosage may need to be maintained for longer than a single month before the index was brought into range. (Thus they were warned about transient responses of the system.) It would be clear if a prescription was too high, subjects were told, because a consistently depressed patient maintained on an overdosage of antidepressant would become manic or manic-depressive. Similarly, a consistently manic or manic-depressive patient maintained on an overdosage of lithium would become depressed.

Each subject in the pair had his or her own joystick and was shown how to use it to answer questions. Moving it from left to right caused all possible answers to the current question to appear on the screen. The joystick was to be moved until the desired response was presented. Pressing a button on the joystick console then registered this response. Subjects were instructed to complete the experiment in silence and not to communicate with their partners in any way.

Patients were presented in random order. To produce initial data on each one, the appropriate A value was assigned, and a random number between 0 and 1 was chosen as the first value of Y. The system was then iterated 100 times to bring it into a stable state determined by the value of A. The next two iterations were used to generate the initial data on the patient. Further iterations producing later patient data were made on the basis of the A values modified by the doctors' treatment decision (i.e. drug prescription) in the manner outlined above.

Patient data, doctors' prescriptions, and the question about treatment efficacy were all presented in the upper "public" portion of the computer screen. Assessments of treatment efficacy were made in the lower "confidential" part. At the top of the screen, the title "Initial Data on Your Patient" appeared on the first month. On later months, this was replaced by "Data on Your Patient One Month Later". Directly below the title, the headings "Happiness Index" and "Treatment Regime" appeared on the left and right, respectively. Beneath these headings, patient data were presented. Thus, for example, when a patient first appeared, the first row of data might have the information "Month 1 = 27 After No Drug." The second row might have the information "Month 2 = 40 After No Drug." Only the two most recent months' data were presented. For example, after the first treatment decision was made and evaluated, the system stepped forward a month such that Month 1 information disappeared, Month 2 information moved into its place, and the new information regarding Month 3 (e.g., "Month 3 = 18 After 8 Mg Lithium") was presented below it.

Below the data on a patient, two questions appeared for the doctor to answer. The first was "Which Drug?" After the doctor had selected "Lithium" or "Antidepressant," the question "What Dosage?" appeared. After the doctor had decided on a dose between 0 and 30 mg, both subjects were presented with the question "What is the probability that this will result in the patient's happiness index being normal (i.e., 29–31) next month?" Each subject then used his or her joystick to select an integer value between 0 and 100% in the lower "confidential" part of the screen. After 10 treatments had been prescribed and assessed, subjects were presented with the next patient.

Results

Confidence. The probability judgments were examined using a judge (doctor vs nurse) by patient difficulty (8 levels) by treatment month (10 levels) repeated-measures analysis of variance (ANOVA). This analysis indicates that doctors (mean confidence 62%) were significantly more confident that their treatments would be effective than were nurses (mean confidence 52%), $F(1, 29) = 8.50$, $p < .01$. For both doctors and nurses, confidence tended to increase over treatment months, $F(9,$

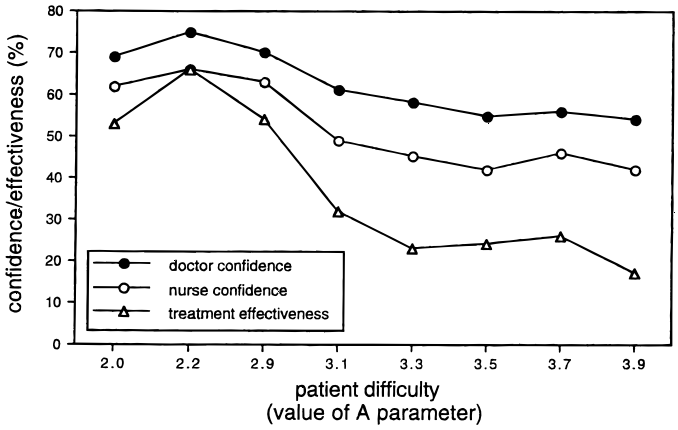

FIG. 1. Doctor and nurse confidence, along with corresponding treatment effectiveness, as a function of patient difficulty (indicated by the value of the parameter *A*) in Experiment 1.

261) = 36.11, $p < .01$, and to decrease as patient difficulty increased, $F(7, 203) = 15.44$, $p < .01$. The increase in confidence over treatment months tended to be greater for doctors than for nurses, $F(9, 261) = 2.21$, $p < .05$. No other effects were statistically significant.

Treatment effectiveness. Mean treatment effectiveness, as measured by the likelihood that the doctor's treatment was successful, was 37%. Both doctors and nurses, then, tended to be overconfident in the sense of overestimating the probability that the treatment would be effective. Doctors were more overconfident than were nurses. A repeated-measures ANOVA of the treatment effectiveness measure indicates that effectiveness tended to increase over treatment months, $F(9, 261) = 15.05$, $p < .01$, and to decrease as patient difficulty increased, $F(7, 203) = 15.77$, $p < .01$. As noted above, these effects were mirrored in the confidence measure, suggesting that subjects were sensitive to their influence on treatment effectiveness. A significant treatment month by patient difficulty interaction, $F(63, 1827) = 1.85$, $p < .01$, suggests that the significant improvement in performance over months was greater for low-difficulty than for high-difficulty patients. No other effects were statistically significant. Figure 1 shows how confidence and treatment effectiveness decreased with increasing patient difficulty. In this and all subsequent experiments, there was no significant improvement in treatment effectiveness over the course of the eight patients that were encountered, as was also found by Harvey (1990b).

Brier score decomposition. The relationship between confidence and treatment effectiveness was assessed in the conventional manner, namely by computing and decomposing Brier scores (Brier, 1950) separately for each subject in the experiment (e.g., see Lichtenstein, Fischhoff, & Phillips, 1982; Yates, 1990, 1994). In the Brier score analyses reported in this paper, the probability judgments were first rounded to the nearest 10%, forming 11 possible probability categories. Table 2 shows the results of the analysis. The overall correspondence between confidence and treatment effectiveness was not particularly impressive in this or subsequent experiments. Indeed, the (mean) Brier scores found in the present experiments are not generally better than that which could be achieved by reporting a uniform confidence value of 50% on all trials, though such a strategy would have an adverse effect on at least one of the component scores, namely resolution. There was a marginally significant overall difference in mean Brier score between doctors and nurses, with doctors giving confidence judgements that correspond less well with treatment effectiveness than those of the nurses.

The source of the difference in Brier scores can be investigated by decomposing the score into theoretically interesting components. We carried out both the Murphy decomposition (Murphy, 1973), which yields calibration and resolution components, and the covariance decomposition (Yates, 1982, 1988), which yields bias, slope, and scatter components. Both decompositions also yield a variance component, which depends entirely on the outcome variable and not on the confidence judgments.

The Murphy decomposition indicates that the difference in Brier scores between doctors and nurses is wholly attributable to the significantly better calibration achieved by nurses than by doctors; no difference was observed for resolution. This result suggests that while the nurses tended to give lower confidence judgements, yielding better calibration, they were no more sensitive to the difference between an effective and ineffective treatment than were the doctors. The difference in calibration between the two conditions is shown in the calibration curve of Fig. 2. It should be noted that these calibration analysis results are very similar to

TABLE 2

Mean Brier Score and Components for Experiment 1, Computed Separately for Each Subject and Then Averaged within Experimental Condition

Measure	Doctor	Nurse	Paired $t(29)$	Significance
Brier	.288	.235	2.03	$p < .06$
Variance	.211	.211	—	—
Calibration	.133	.080	2.10	$p < .05$
Resolution	.0552	.0555	0.05	n.s.
Bias	.262	.159	2.82	$p < .01$
Slope	.179	.225	1.96	$p < .06$
Scatter	.0371	.0546	2.72	$p < .02$

Note. Also shown are the *t* test value and observed significance level comparing doctors and nurses on each measure.

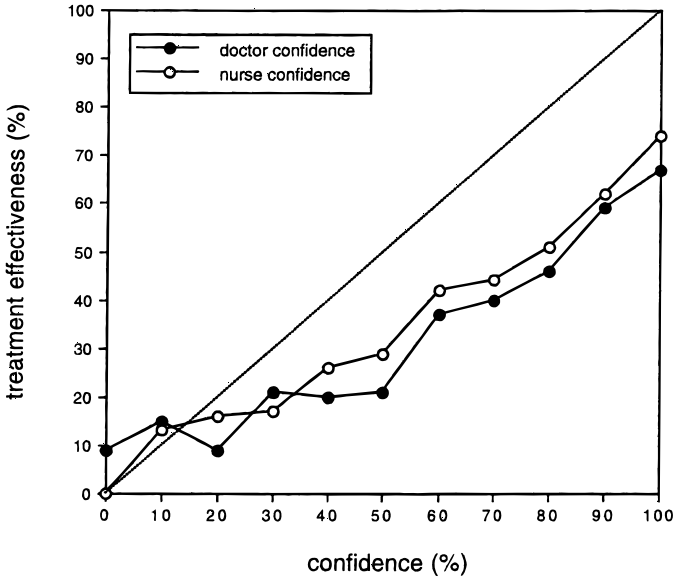

FIG. 2. Calibration curves for doctors and nurses in Experiment 1.

those obtained by Harvey (1990b) in an experiment similar to the present one but in which no nurses were present. This suggests that the mere presence of nurses (i.e., observers) had little effect on the way that doctors evaluated the effectiveness of their treatment decisions.

The covariance decomposition results yield similar conclusions. A large difference in bias (mean overconfidence) indicates that nurses tended to give lower confidence assessments, but the marginal difference in slope suggests that the two groups did not differ greatly in their ability to distinguish successful from unsuccessful treatments. The difference in scatter indicates that the nurses' confidence assessments were "noisier" than those of the doctors, that is, included more variance unrelated to treatment effectiveness.

Discussion

The first experiment demonstrated significant actor-observer differences in confidence, with the doctors expressing greater confidence in their treatment decisions than that expressed by the observing nurses. There are two possible explanations for the relatively lower confidence expressed by the nurses. First, simple difference of opinion regarding the optimal treatment might have led the nurses to feel less confident than the doctors. That is, because doctors are allowed to choose their preferred treatment, any disagreement between the doctor and the nurse regarding the most effective treatment would result in lower confidence for nurses in the doctors' decisions. Second, the difference might be attributable to the fact that the doctors but not the nurses had control of the system (i.e., the decision process). This perception of control on the part of doctors might have enhanced their feeling of confidence relative to that of the nurses.

EXPERIMENT 2

A second experiment was conducted to disentangle the two possible interpretations of the results of Experiment 1. As in the first experiment, the doctor prescribed a treatment on each trial which was then evaluated by both the doctor and the nurse. Following this, however, the nurse then gave his or her own treatment recommendation. Both doctor and nurse then assessed their confidence in the nurses' recommended treatment. Only the doctor's treatment was implemented on each trial. If the doctor–nurse confidence difference observed in the previous experiment is due to simple difference of opinion regarding treatment, then there should be symmetric effects on confidence for the doctors' and nurses' decisions: Doctors should be more confident than nurses in the doctors' decisions and less confident than nurses in the nurses' decisions. On the other hand, if the effect depends on perceived control over the system, then we should expect to find a difference only for the doctors' decisions as it is only their decisions that are actually implemented on each trial.

Method

Subjects. A new group of 60 undergraduate volunteers aged between 18 and 30 years served as subjects, yielding 30 doctor–nurse pairs.

Procedure. Each trial of the experiment proceeded as follows. First, the doctor made a treatment decision on the basis of the information provided about the patient. Both doctor and nurse then made (independent) judgments of the probability that the doctor's treatment would be effective. Following this, the nurse was asked to give his or her own treatment recommendation for the patient. Both the doctor and the nurse then gave probability judgments that the nurse's recommended treatment would be effective if it were to be implemented. At this point the doctor's treatment was implemented, feedback regarding the results of this treatment was provided, and the next trial (monthly treatment) commenced. In all other respects the second experiment was identical to the first.

Results

Confidence. The probability judgments (in this and in all subsequent experiments) were analyzed using a treatment (doctor vs nurse) by judge (doctor vs nurse) by patient difficulty (8 levels) by treatment month (10 levels) repeated-measures ANOVA. A significant treatment by judge interaction indicates that differences in

expressed confidence between doctors and nurses depended on whether the doctor's or nurse's treatment was being assessed, $F(1, 29) = 59.02$, $p < .01$. When evaluating the doctors' treatments, doctors (mean confidence 60%) and nurses (mean confidence 58%) expressed essentially equal confidence, simple effects $F(1, 29) < 1$. When assessing the nurses' treatments, in contrast, nurses (mean confidence 65%) were significantly more confident than were doctors (mean confidence 53%), simple effects $F(1, 29) = 14.81$, $p < .01$.

The analysis also revealed effects of patient difficulty and treatment month similar to those found in the previous experiment. For both doctors and nurses, confidence tended to increase over treatment months, $F(9, 261) = 61.24$, $p < .01$, and to decrease as patient difficulty increased, $F(7, 203) = 7.28$, $p < .01$. Furthermore, confidence tended to increase over treatment months more quickly for low-difficulty patients than for high-difficulty patients, $F(63, 1827) = 2.10$, $p < .01$. Finally there was an apparently anomalous treatment by difficulty interaction, $F(7, 203) = 2.71$, $p < .01$, which is not easily interpretable and did not occur in any of the subsequent experiments. No other effects were statistically significant.

Treatment effectiveness. On average, the doctors' treatments were effective in 39% of the cases. Treatment effectiveness for nurses, computed by determining the expected outcome had the treatment been implemented, was 38%. The difference in treatment effectiveness between doctors and nurses is nonsignificant, $F(1, 29) < 1$. Thus both doctors and nurses were overconfident in their assessments of both doctor and nurse treatments, and nurses were more overconfident than doctors in assessing the nurses' treatments.

Again, as reflected in the confidence judgments, treatment effectiveness increased with treatment month, $F(9, 261) = 15.11$, $p < .01$, and decreased as patient difficulty increased, $F(7, 203) = 14.33$, $p < .01$. Furthermore, treatment effectiveness—as well as confidence—tended to increase over treatment months more quickly for low-difficulty patients than for high-difficulty patients $F(63, 1827) = 1.55$, $p < .01$. The only other statistically significant effect was an uninterpretable treatment by treatment month interaction, $F(9, 261) = 1.94$, $p < .05$, which did not occur in any of the subsequent experiments. Because similar results were obtained in the following two experiments, Figure 3 combines the data from Experiments 2, 3, and 4 and displays confidence and treatment effectiveness as a function of patient difficulty separately for the doctors' and nurses' treatments.

Brier score decomposition. For each subject, Brier scores were computed and decomposed separately for the nurses' treatments and for the doctors' treatments. The results appear in Table 3. Reflecting the overconfidence measure, doctors and nurses achieved essentially equal Brier scores when judging the doctors' treatments but doctors achieved better scores than the nurses when assessing the nurses' decisions. The Murphy decomposition of the Brier scores indicated that the difference in the latter case was again completely attributable to differences in calibration and not in resolution. This was also reflected in the covariance decomposition, which indicated differences in bias but not slope in judgments of the nurses' treatments. In this and subsequent experiments, there was a pattern of greater scatter for judgments of the experimental partner's treatments than for judgments of one's own treatments. We return to this observation in the General Discussion. Figure 4 displays calibration curves; once again, the data from Experiments 2, 3, and 4 have been combined because their results are quite similar.

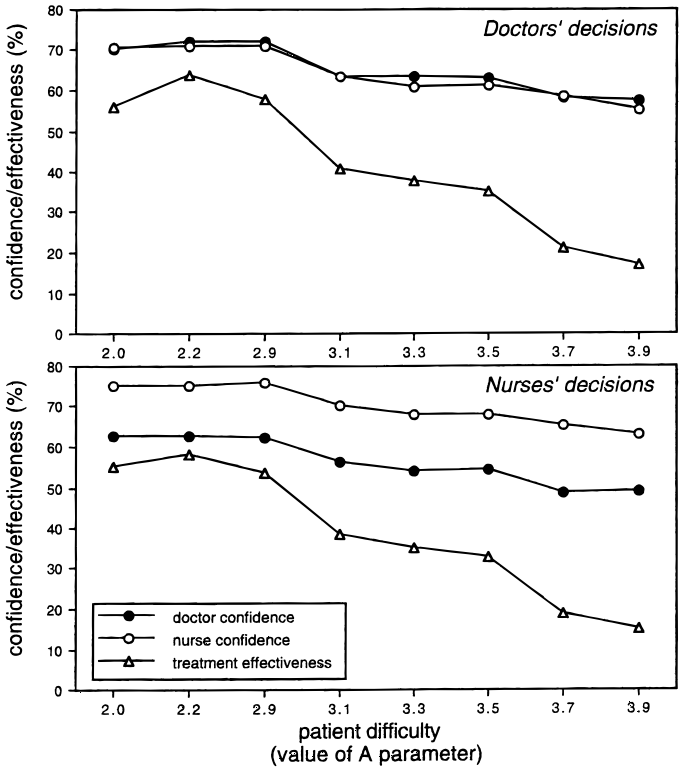

FIG. 3. Doctor and nurse confidence, along with corresponding treatment effectiveness, as a function of patient difficulty (indicated by the value of the parameter *A*). The top panel represents assessments of the doctors' treatment recommendations; the bottom represents assessments of the nurses' treatment recommendations. Data from Experiments 2, 3, and 4 have been combined in this figure.

Discussion

In the first experiment, doctors expressed greater confidence in their treatment recommendations than

TABLE 3

Mean Brier Score and Components for Experiment 2, Computed Separately for Each Subject and Then Averaged within Experimental Condition

Measure	Doctor	Nurse	Paired $t(29)$	Significance
Doctor treatment				
Brier	.260	.263	0.18	n.s.
Variance	.210	.210	—	—
Calibration	.0966	.102	0.29	n.s.
Resolution	.0465	.0484	0.33	n.s.
Bias	.212	.199	0.44	n.s.
Slope	.125	.139	0.85	n.s.
Scatter	.0369	.0455	1.93	$p < .07$
Nurse treatment				
Brier	.245	.310	2.59	$p < .02$
Variance	.208	.208	—	—
Calibration	.0841	.145	2.41	$p < .03$
Resolution	.0469	.0414	1.18	n.s.
Bias	.151	.281	3.98	$p < .01$
Slope	.137	.111	1.76	n.s.
Scatter	.0457	.0316	2.40	$p < .03$

Note. Also shown are the t test value and observed significance level comparing doctors' and nurses' judgments on each measure, listed separately for doctors' and nurses' treatments.

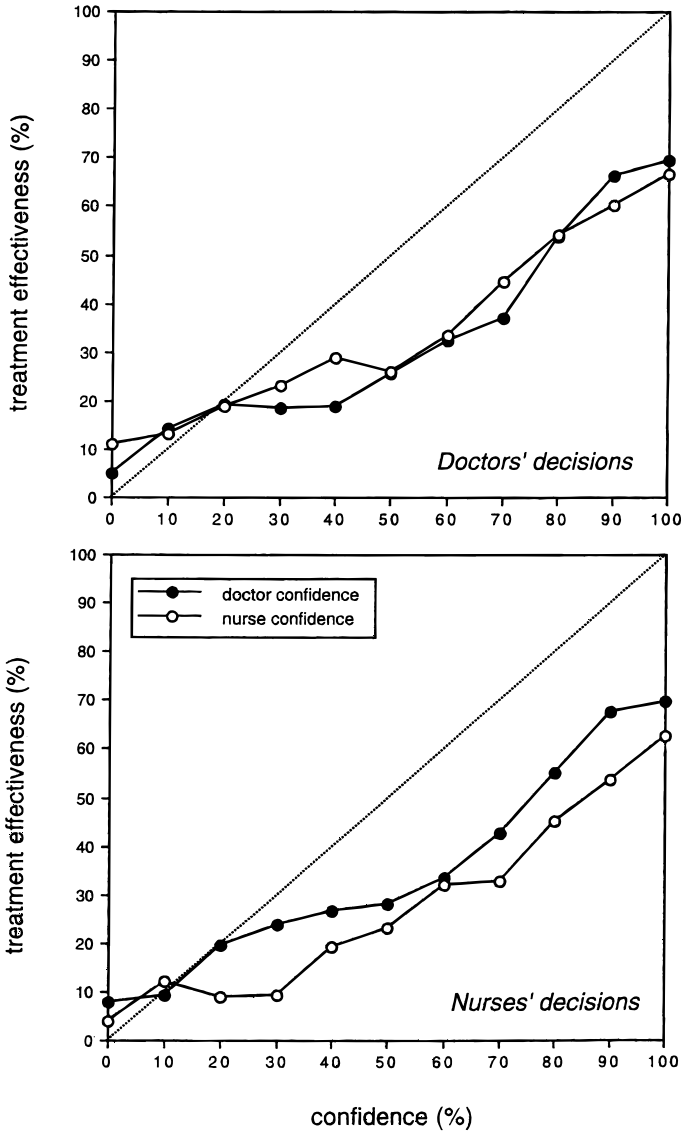

FIG. 4. Calibration curves for doctors and nurses, constructed by combining the data from Experiments 2, 3, and 4. The top panel represents assessments of the doctors' treatment recommendations; the bottom represents assessments of the nurses' treatment recommendations.

that expressed by nurses. This effect was unexpectedly eliminated in the second experiment, however, when nurses were given the opportunity to make their own, alternative treatment recommendations. The effect observed in the first experiment for the doctors' treatments did hold, however, for the nurses' treatments in the second experiment (i.e., nurses were more confident than were doctors). One possible interpretation of these findings is that they are attributable to a "second-guessing" effect: The doctors, now knowing that every treatment recommendation they make will be immediately followed by the nurses' own treatment recommendation, may be made more sensitive as a result to the existence or validity of alternative treatment recommendations. The doctors' confidence in their treatment recommendations might have decreased as a consequence of this salient second guessing on the part of the nurses. The second-guessing effect might have also increased the nurses' confidence in the doctors' recommendations, in that they could not give skeptical assessments without subsequently offering a better suggestion regarding treatment. Thus, on this interpretation, the second-guessing effect should move doctors' and nurses' confidence in the doctors' treatments toward each other relative to the first experiment.

This interpretation suggests that the order in which the doctor and nurse give their treatment recommendations is crucial. That is, doctors and nurses were equally confident in the doctors' recommendation because both knew that the nurse would subsequently be giving his or her own recommendation. In short, the doctors' recommendations (but not the nurses') were being subjected to second guessing. On this account, the effect should be reversed if the nurses' recommendations are given before the doctors'.

An alternative interpretation is that the difference in judgments of doctors' and nurses' recommendations arises from the fact that the doctors' recommendations are implemented and the nurses' are not. For example, the feedback received regarding the doctors' recommendations may drive assessments regarding these recommendations into general agreement, while the lack of feedback regarding the nurses' recommendations may

allow the doctors and nurses to continue to disagree when assessing those recommendations. (Note that this interpretation must be supplemented in some way to account for the discrepancy between doctors' and nurses' assessments in the first experiment. One possibility is that the generation of alternatives on the part of nurses in the second experiment makes both doctors and nurses more aware of alternative treatments, and that this awareness is necessary—along with the feedback—to bring the doctors' and nurses' confidence judgments into agreement.) On this alternative interpretation, the order in which the doctors' and nurses' recommendations are given should have no effect.

EXPERIMENT 3

The third experiment was conducted to test between the two alternative interpretations outlined above. This experiment was identical to the previous one, except that the nurses gave their recommendations before the doctors.

Method

Subjects. A new group of 60 undergraduate volunteers aged between 18 and 30 years served as subjects, yielding 30 doctor–nurse pairs.

Procedure. The experimental design and procedure were identical to that of the previous experiment, the only difference being the reversal of the order in which the doctors and nurses gave their treatment recommendations. Thus, on a given trial, the nurse first gave a treatment recommendation, and both nurse and doctor assessed their confidence in this recommendation, then the doctor gave a treatment recommendation and both subjects again made confidence judgments. As in the previous experiment, only the doctors' recommendations were implemented. Feedback, then, was available regarding the doctors' but not the nurses' treatment recommendations.

Results

Confidence. The results of this experiment were essentially identical to those of Experiment 2. A significant treatment by judge interaction indicates that differences in expressed confidence between doctors and nurses depended on whether the doctor's or nurse's treatment was being assessed, $F(1, 29) = 43.11$, $p < .01$. When assessing the nurses' treatments, nurses (mean confidence 72%) were significantly more confident than were doctors (mean confidence 59%), simple effects $F(1, 29) = 25.51$, $p < .01$. When evaluating the doctors' treatments, in contrast, doctors (mean confidence 69%) and nurses (mean confidence 67%) expressed essentially equal confidence, simple effects $F(1, 29) = 1.50$, n.s. Reversing the order in which the treatment recommendations were made, then, had no effect on the general pattern of results.

The analysis also revealed effects of patient difficulty and treatment month similar to those found in the previous experiments. For both doctors and nurses, confidence tended to increase over treatment months, $F(9, 261) = 35.21$, $p < .01$, and to decrease as patient difficulty increased, $F(7, 203) = 13.96$, $p < .01$. Furthermore, confidence tended to increase over treatment months more quickly for low-difficulty patients than for high-difficulty patients, $F(63, 1827) = 2.66$, $p < .01$. No other effects were statistically significant.

Treatment effectiveness. On average, the doctors' treatments were effective in 43% of the cases. Treatment effectiveness for nurses, computed by determining the expected outcome had the treatment been implemented, was 38%. This difference, in contrast to the previous experiment, is statistically significant, $F(1, 29) = 6.69$, $p < .05$. Both doctors and nurses were overconfident in their assessments of both doctor and nurse treatments; nurses were more overconfident than doctors when assessing the nurses' treatments but not when assessing the doctors' treatments.

Again, as reflected in the confidence judgments, treatment effectiveness increased with treatment month, $F(9, 261) = 19.15$, $p < .01$, and decreased as patient difficulty increased, $F(7, 203) = 19.58$, $p < .01$. Furthermore, treatment effectiveness—as well as confidence—tended to increase over treatment months more quickly for low-difficulty patients than for high-difficulty patients, $F(63, 1827) = 2.23$, $p < .01$. The only other statistically significant effect was a treatment by treatment month by patient difficulty interaction $F(9, 261) = 1.94$, $p < .05$, in which the improvement of nurses' treatments over treatment months was slower than that of doctors' for high-difficulty (but not low-difficulty) patients. This effect did not occur in any of the other experiments.

Brier score decomposition. For each subject, Brier scores were computed and decomposed separately for the nurses' treatments and for the doctors' treatments. The results of the Brier score analysis, which appear in Table 4, follow the same pattern as that observed in the previous experiment. Doctors and nurses achieved essentially equal Brier scores when judging the doctors' treatments but doctors achieved better scores than the nurses when assessing the nurses' decisions. Decomposition of the Brier scores indicated that the difference in the latter case was again completely attributable to

TABLE 4

Mean Brier Score and Components for Experiment 3, Computed Separately for Each Subject and Then Averaged within Experimental Condition

Measure	Doctor	Nurse	Paired $t(29)$	Significance
Doctor treatment				
Brier	.288	.288	0.04	n.s.
Variance	.221	.221	—	—
Calibration	.124	.120	0.39	n.s.
Resolution	.0565	.0528	0.67	n.s.
Bias	.271	.252	1.27	n.s.
Slope	.155	.150	0.34	n.s.
Scatter	.0339	.0455	2.94	$p < .01$
Nurse treatment				
Brier	.255	.321	3.31	$p < .01$
Variance	.208	.208	—	—
Calibration	.0989	.166	3.81	$p < .01$
Resolution	.0518	.0537	0.45	n.s.
Bias	.218	.339	4.94	$p < .01$
Slope	.195	.161	1.67	n.s.
Scatter	.0474	.0368	2.80	$p < .01$

Note. Also shown are the t test value and observed significance level comparing doctors' and nurses' judgments on each measure, listed separately for doctors' and nurses' treatments.

differences in calibration (or bias) and not in resolution (or slope).

Discussion

The essentially identical results of Experiments 2 and 3 indicate that the order in which doctors' and nurses' recommendations are elicited has no substantial effect on judged confidence or its relationship to the treatment effectiveness measure. Such a result would appear to discredit the "second-guessing effect" interpretation of the asymmetry in judgments of doctors' and nurses' recommendations, and to implicate the asymmetry of feedback instead as the source of the results. The suggestion, then, is that the availability of feedback regarding the doctors'—but not the nurses'—recommendations, coupled with the increased salience of alternative possible treatments caused by asking the nurses to give their own recommendations, drives the doctors' and nurses' confidence judgments regarding the doctors' treatment recommendations into agreement. Because no feedback is given regarding the nurses' recommendations, in contrast, doctors and nurses can continue to disagree in their assessments of these recommendations.

With this possibility in mind, the data of Experiments 2 and 3 were re-analyzed by substituting patient number (i.e., the order in which the eight patients were encountered in the experiment) for patient difficulty in the repeated-measures ANOVA. (Recall that each subject was presented with the eight patients in a randomized order, such that, over subjects, patient difficulty is expected to be approximately equal for all eight positions in the sequence.) If the feedback regarding the doctors' recommendations plays a crucial role, we might expect to find symmetric differences very early in the sequence (e.g., for the first patient), such that doctors are more confident than nurses in doctor recommendations and vice versa for the nurse recommendations. As more patients are encountered, however, the impact of the feedback should increase, yielding increasingly asymmetric differences between doctor and nurse confidence for the two recommendations.

Indeed, the treatment by judge by patient number interaction suggested by this argument was significant in both experiments, $F(7, 203) = 2.79$ in Experiment 2 and 2.81 in Experiment 3, both ps $< .01$. This effect can be seen in Fig. 5, which combines the data from the two experiments. As can be seen in the figure, the interaction arises because the doctor–nurse difference has a roughly constant negative value for nurses' recommendations (indicating greater confidence by nurses than doctors), while the doctor–nurse difference for the doctors' recommendations is positive (indicating greater confidence by doctors than nurses) only very early in the patient sequence, and is quickly driven to zero. (Note that this result cannot be attributed to changes in treatment effectiveness, which remained constant across the eight-patient sequence.) Such a result is consistent with the idea that both the doctor and the nurse begin with symmetrically greater confidence in their own than in the other's recommendations, but subsequent feedback regarding the doctor's recommendations forces judgments regarding these recommendations into agreement.

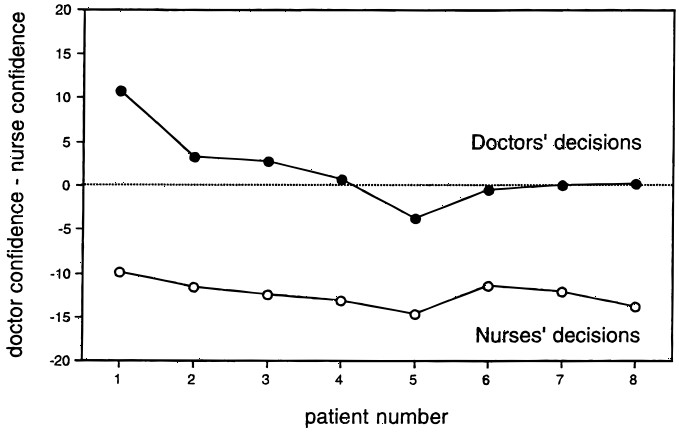

FIG. 5. Average difference between doctor confidence and nurse confidence as a function of patient number (i.e., the patient's position in the eight-patient experimental sequence), plotted separately for doctors' and nurses' treatment recommendations. Data from Experiments 2 and 3 have been combined in this figure.

The obvious way to test this interpretation is to give feedback regarding the nurses' recommendations in addition to that given regarding the doctors' recommendations. If the interpretation is correct, we would expect the resulting differences between doctors and nurses in their confidence judgments to be symmetric and less pronounced (or even eliminated) relative to the differences found in the absence of feedback. This possibility is tested in Experiment 5. Before moving on to that experiment, however, we felt it was necessary to rule out an artifactual explanation of the doctor–nurse differences observed in the previous experiments.

EXPERIMENT 4

It is possible that the differences observed in the previous experiments between doctors' and nurses' judgments are due not to manipulations of the availability of feedback or the salience of alternative possible treatments, but rather to an unintentional social psychological effect arising from the role designations used in these experiments. Specifically, the assignment of subject pairs to doctor and nurse roles may have somehow led subjects to view the doctor subjects as more knowledgeable or competent in the task than the nurse subjects. However unlikely, we thought it prudent to rule out this possibility by conducting an experiment in which the nurses were now referred to as "consultant psychiatrists."

Method

Subjects. A new group of 60 undergraduate volunteers aged between 18 and 30 years served as subjects, yielding 30 doctor–nurse pairs.

Procedure. The experimental design and procedure were identical to that of Experiment 2, with doctors giving their recommendations before the nurses, the only difference being that the nurses were now referred to as experienced "consulting psychiatrists" observing the actions of the more junior doctors. For consistency, these subjects will continue to be referred to as "nurses" in this paper. As in the previous experiment, only the doctors' recommendations were implemented, so that feedback was available regarding the doctors' but not the nurses' (consulting psychiatrists') treatment recommendations.

Results

Confidence. The results did not differ in any substantial way from those of Experiment 2, suggesting that the designation of the observer subject as a "nurse" in previous experiments did not, in and of itself, affect the results. Because the results are so similar to previous experiments they are discussed only briefly here. When evaluating the doctors' treatments, doctors (mean confidence 66%) and nurses (mean confidence 66%) expressed essentially equal confidence, simple effects $F(1, 29) < 1$. When assessing the nurses' treatments, in contrast, nurses (mean confidence 74%) were significantly more confident than were doctors (mean confidence 57%), simple effects $F(1, 29) = 58.19$, $p < .01$. Effects of patient difficulty and treatment month were similar to those found in previous experiments.

Treatment effectiveness. On average, the doctors' treatments were effective in 41% of the cases. Treatment effectiveness for nurses, which did not differ significantly from that of doctors, was 39%. Thus both doctors and nurses were overconfident in their assessments of both doctor and nurse treatments, and nurses were more overconfident than doctors in assessing the nurses' treatments. Treatment effectiveness, like confidence, was affected by patient difficulty and treatment month as in previous experiments.

Brier score decomposition. Analysis once again showed that doctors and nurses achieved essentially equal Brier scores when judging the doctors' treatments but doctors achieved better scores than the nurses when assessing the nurses' decisions. In contrast to the previous experiments, however, decomposition of the Brier scores indicated significant differences in resolution (and slope) as well as calibration (and bias). It is unclear why discrimination ability (measured by resolution or slope) was affected in only this experiment.

Discussion

The artifactual interpretation regarding the importance of designating the observer subjects as nurses appears to be discredited. Having dismissed this alternative, we can move on to investigate the influence of feedback on the pattern of results obtained in the earlier experiments.

EXPERIMENT 5

In the final experiment, feedback was provided for both the doctors' and the nurses' recommendations. Only the doctors' treatment was actually implemented on each trial, but subjects were informed on each trial what would have happened had the nurses' recommendation been implemented instead. This design allows us to disentangle the influences of feedback and control over the system. The interpretation offered earlier implicates feedback as the crucial variable, and predicts symmetric effects on judgments of the doctors' and

nurses' recommendations. If perceived control is the important variable, in contrast, we would expect continued asymmetries in the judgments as only the doctors have control over the system (i.e., the course of treatment actually implemented on each trial).

Method

Subjects. A new group of 60 undergraduate volunteers aged between 18 and 30 years served as subjects, yielding 30 doctor–nurse pairs.

Procedure. With the exception of the additional feedback, the experimental design and procedure were essentially identical to that of Experiment 2. The doctors first gave their recommendations, which were assessed by both subjects; then the nurses gave their recommendations, which were also assessed by both subjects. After this process was completed, in contrast to the previous experiments, subjects were presented with feedback regarding both the outcome of the doctors' treatment recommendation and the expected outcome of the nurses' recommendation, had it been implemented. Subjects were told that the effectiveness of the nurse's treatment could be established on the basis of knowledge of how various dosage levels had affected patients similar to the one undergoing treatment. Once this feedback was presented, subjects moved on to the next trial. In all other respects, the procedure was identical to that of Experiment 2.

Results

Confidence. In contrast to the previous three experiments, symmetric doctor-nurse differences in confidence were obtained for the doctors' and nurses' recommendations. A significant treatment by judge interaction indicates that differences in expressed confidence between doctors and nurses depended on whether the doctor's or nurse's treatment was being assessed, $F(1, 29) = 68.98$, $p < .01$. When evaluating the doctors' treatments, doctors (mean confidence 65%) were more confident than nurses (mean confidence 57%), simple effects $F(1, 29) = 8.66$, $p < .01$. When assessing the nurses' treatments, in contrast, nurses (mean confidence 63%) were significantly more confident than were doctors (mean confidence 55%), simple effects $F(1, 29) = 13.95$, $p < .01$. Once feedback was given for both recommendations, then, the asymmetry in the judgments found in earlier studies was eliminated: Both doctors and nurses expressed greater confidence in their own recommendation than they did in that of their partner.

Analysis of the confidence judgments once again revealed effects of patient difficulty and treatment month. For both doctors and nurses, confidence tended to increase over treatment months, $F(9, 261) = 10.84$, $p < .01$, and to decrease as patient difficulty increased, $F(7, 203) = 7.90$, $p < .01$. No other effects were statistically significant.

Treatment effectiveness. On average, the doctors' treatments were effective in 26% of the cases. Treatment effectiveness for nurses, based on the expected outcome had the treatment been implemented, was 25%. This difference is not significant, $F(1, 29) < 1$. Note that the rate of treatment effectiveness was lower than in the previous experiments, perhaps because subjects had to divide their attention between the two sources of feedback. Both doctors and nurses were overconfident in their assessments of both doctor and nurse treatments, with both groups being more overconfident in their own recommendations than in those of their partners.

As in previous experiments, treatment effectiveness increased with treatment month, $F(9, 261) = 8.88$, $p < .01$, and decreased as patient difficulty increased, $F(7, 203) = 17.83$, $p < .01$. Furthermore, treatment effectiveness—but not confidence—tended to increase over treatment months more quickly for low-difficulty patients than for high-difficulty patients, $F(63, 1827) = 1.46$, $p < .05$. No other effects were significant.

Brier score decomposition. The results of the Brier score analysis appear in Table 5. As in the first experiment, and in contrast to Experiments 2 to 4, nurses

TABLE 5

Mean Brier Score and Components for Experiment 5, Computed Separately for Each Subject and Then Averaged within Experimental Condition

Measure	Doctor	Nurse	Paired $t(29)$	Significance
Doctor treatment				
Brier	.358	.291	3.02	$p < .01$
Variance	.170	.170	—	—
Calibration	.225	.158	3.31	$p < .01$
Resolution	.0371	.0368	0.09	n.s.
Bias	.402	.322	2.93	$p < .01$
Slope	.121	.125	0.21	n.s.
Scatter	.0456	.0453	0.08	n.s.
Nurse treatment				
Brier	.300	.338	1.95	$p < .07$
Variance	.173	.173	—	—
Calibration	.169	.199	1.70	n.s.
Resolution	.0419	.0349	1.81	n.s.
Bias	.308	.386	3.67	$p < .01$
Slope	.122	.113	0.49	n.s.
Scatter	.0524	.0349	4.25	$p < .01$

Note. Also shown are the t test value and observed significance level comparing doctors' and nurses' judgments on each measure, listed separately for doctors' and nurses' treatments.

achieved better scores than the doctors when assessing the doctors' decisions. Decomposition of these Brier scores indicated that the superiority of the nurses' judgments was entirely attributable to differences in calibration rather than resolution. For the nurses' decisions, there were only marginal differences in Brier scores, which were in the direction of better scores for doctors than for nurses. Again, the superiority of the doctors' judgments appeared to be attributable to differences in calibration (or bias) rather than resolution (or slope). In both cases, observers tended to give less overconfident judgments than actors.

Discussion

The final experiment clearly implicates feedback as an important factor influencing actor–observer differences in judgment. The asymmetric doctor–nurse differences in confidence observed in Experiments 2 to 4 were rendered symmetric once feedback regarding nurses' as well as doctors' recommendations was provided. Under these conditions, doctors and nurses exhibited comparable tendencies toward exhibiting greater confidence in their own than in their experimental partner's treatment recommendations. This observation suggests that having (or perceiving) control over the system (i.e., over which treatment is actually implemented on a given trial) has little influence on confidence, at least in the task we studied. As long as the subject is informed about the expected outcome of his or her recommended treatment, it seems to matter little (in terms of judged confidence) whether that recommendation is actually implemented or not.

We suggested earlier that the feedback regarding the doctors' recommendations given in Experiments 2 and 3 was responsible for driving the doctors' and nurses' confidence judgments regarding these recommendations into greater agreement as more and more patients were encountered (see Fig. 5). This argument might be taken to suggest that, in the final experiment, a comparable effect on doctor-nurse agreement should be found for both doctors' and nurses' recommendations because feedback was given for both. The relevant treatment by judge by patient number interaction is indeed significant, $F(7, 203) = 2.25$, $p < .05$. Examination of the mean doctor-nurse confidence differences shows that for assessments of both recommendations, the difference in confidence between doctors and nurses was greater for the first four patients ($M = +9.3\%$ for doctor recommendations and $M = -9.0\%$ for the nurse recommendations) than for the last four patients ($M = +6.7\%$ for doctor recommendations and $M = -6.7\%$ for the nurse recommendations) seen in the experiment. Thus, although it did not drive the doctor-nurse difference to zero as in the earlier experiments, the presence of feedback did lead to smaller differences as more and more patients were seen.

GENERAL DISCUSSION

Two simple hypotheses regarding actor–observer differences in judging the probability of treatment effectiveness might be used to predict the results of our experiments. (1) According to the *control* hypothesis, people tend to overestimate their ability to control the output of complex systems such as the one examined in this paper. By this account, we would expect doctors to be overconfident in their treatment recommendations, while nurses would be expected to give more realistic assessments of the doctors' recommendations. Because the nurses' recommendations, in contrast, were not implemented, we would not expect to find a similar effect in assessments of their recommendations. (2) According to the *disagreement* hypothesis, actor-observer differences arise from differences in opinion regarding the appropriate course of action. On this account, we would expect doctors to be more confident than nurses in the doctors' recommendations and nurses to be more confident than doctors in the nurses' recommendations.

The results from the five experiments presented here are not consistent with either of these simple hypotheses. Instead, a more complex pattern of results was observed. While doctors were indeed more confident than nurses in the doctors' treatment recommendations in the first experiment, this effect was eliminated in Experiments 2 to 4 when the nurses were given the opportunity to offer their own, alternative recommendations. It is not entirely clear whether this change was due to lowered confidence on the part of doctors when faced with the nurses' competing recommendations, or due instead to increased confidence in the doctors' recommendations on the part of nurses when obliged to justify, in a sense, their own skepticism of the doctors' recommendations by coming up with a better idea of their own. Examination of mean confidence for the first two patients in Experiments 2 and 3 suggests that both of these phenomena played a role: Between the first and second patient, the doctors' confidence in the doctors' recommendations decreased while the nurses' confidence in these recommendations increased.

The present results are also apparently inconsistent with the studies by Koehler (1994) showing that subjects who generate their own hypotheses express lower confidence in their truth than do observers presented with the same hypotheses for evaluation. The differences between the tasks investigated by Koehler (1994) and that used in the present studies, however, are so numerous as to make direct comparison of the results

quite difficult. Koehler and Harvey (in press) present the results of an initial attempt to identify the source of these apparently contradictory findings.

Although the requirement that nurses give their own recommendations eliminated the previously-observed difference in confidence between doctors and nurses in the doctors' recommendations, such a difference reemerged in assessments of the nurses' recommendations. A consistent pattern was found in Experiments 2 to 4 of greater confidence by nurses than doctors in the nurses' recommendations, but equal confidence in the doctors' recommendations. Experiment 5 showed that this asymmetry is due, not to differences in control (i.e., treatment implementation) between doctors and nurses, but rather to the absence of feedback regarding the nurses' treatment recommendations. Once feedback was given for both doctors' and nurses' recommendations, the asymmetry in the probability judgments was eliminated.

The importance of feedback was highlighted by an analysis of how differences in confidence between doctors and nurses changed over the eight-patient experimental sequence. As shown in Fig. 5, doctors were initially more confident than nurses in the doctors' recommendations, but this difference was eliminated after the first couple of patients encountered. We assume that this reduction in disagreement was due to the feedback received regarding the outcome of the doctors' recommendations. Because no such feedback was given in these experiments regarding the nurses' recommendations, in contrast, the greater confidence of nurses than of doctors in the nurses' recommendations remained relatively constant throughout the experiment.

Rather than driving confidence differences to zero for the nurses' recommendations as well as for the doctors' recommendations, however, the effect in Experiment 5 of giving feedback for both recommendations on each trial was instead to reinstate the initial bias in favor of one's own recommendation over that of one's partner. While subsequent analysis showed that in fact this bias was larger at the beginning of the experiment than at the end, suggesting that the feedback had some influence in reducing disagreement between the doctors and nurses, the results clearly indicate that feedback is not sufficient to eliminate actor-observer differences in confidence.

Why is it that the doctors were more confident than the nurses in the doctors' recommendations when no alternative was given by nurses (Experiment 1) and when the nurses did give an alternative recommendation for which feedback was given (Experiment 5), but not when nurses gave alternative recommendations for which no feedback was given (Experiments 2 to 4)? In the absence of feedback regarding the nurses' alternative recommendations, doctors faced the nagging possibility that the nurses' recommendations were better than their own. When coupled with feedback indicating that their own treatments were far from perfectly effective, the doubt raised by this possibility might have led to a decrease in initial confidence over the first few patients. The possibility that the nurses' recommendations were superior was discredited by the provision of feedback regarding their recommendations in Experiment 5, which showed approximately equal effectiveness for doctors' and nurses' treatments. Freed from the doubt raised by this possibility, doctors were once again more confident than nurses in the doctors' treatments.

Two comments are in order regarding the nature and impact of the feedback provided in our experiments, as it appears to play a central role in the results. First, note that the feedback provided in the "dynamic" task we used is quite different from the type of feedback that might be made available in the usual "static" tasks (e.g., general knowledge questions) found in many studies of confidence in judgment. Indeed, outcome feedback may be more informative in a dynamic task than in a static one, in that it has more relevance for the subsequent judgment that is to be made. For example, discovering that one's initial treatment recommendation was not effective in bringing the desired patient index into range serves (at least) two functions for the next judgment: (a) it conveys some sense of the appropriateness of one's previous confidence assessment; and (b) it may also revise one's estimate of the patient's responsiveness to treatment. This latter factor, which reflects the influence of feedback on learning about the dynamic system itself, is arguably absent from static judgment tasks. (The closest analog is the subject's sense of the overall task difficulty level.)

A second comment regarding the influence of feedback in our experiments is that although it appears, under circumstances outlined above, to have driven the confidence assessments of doctors and nurses into closer agreement with each other, it did not drive the assessments into agreement with the actual effectiveness of the treatment. That is, the systematic outcome feedback regarding the doctors' treatment recommendations did not appear to eliminate or even diminish overconfidence in the effectiveness of the recommended treatment over the course of the experiment. For example, doctors were no less overconfident, on average over Experiments 2 to 4, when treating the final four patients they encountered (mean overconfidence = 25%) than they were for the first four (mean overconfidence = 23%). This may be somewhat surprising, given the immediacy and clarity of the provided feedback. Under

less idealized conditions in which there is a lag between the treatment decision and subsequent outcome information, performance in dynamic systems generally deteriorates (e.g., Diehl & Sterman, 1995). Even under relatively ideal conditions, then, subjects were consistently overconfident despite the systematic feedback they received.

In fact, in all conditions of all the experiments we report, subjects were substantially overconfident, in the sense of overestimating the probability that the treatment would be effective. This overestimation, which Yates (1990) refers to as *overconfidence in one's actions*, is arguably a qualitatively different phenomenon from that observed in many calibration studies using general knowledge items, which Yates (1990) refers to as *overconfidence in one's judgment*. It is notable in this regard, however, that perceived control did not appear to play a major role in our results. One likely contributor to the levels of overconfidence observed in our experiments is the overall difficulty of the task we used. Recommended treatments were effective, on average, less than 40 percent of the time. On this basis alone, the well-known effect of difficulty on overconfidence in calibration studies (the difficulty or "hard–easy" effect; e.g., Griffin & Tversky, 1992; Juslin, Olsson, & Björkman, in press; Keren, 1991; Lichtenstein *et al.*, 1982) would lead us to expect considerable overconfidence.

Because of the presence of general overconfidence, the experimental conditions yielding the highest levels of confidence also yielded the most overconfidence. Thus, it can be said that nurses gave better (i.e., less overconfident) assessments of doctor treatments than did doctors in the first and final experiment, and that doctors gave better assessments of nurse treatments than did nurses in Experiments 2 to 4. Can we be more specific about how the confidence assessments of observers (i.e., the person not responsible for the treatment under evaluation) were better in these cases? The Murphy decomposition of the Brier scores indicates that the judgments in these cases were better calibrated but did not have greater resolution. In other words, observers under these circumstances tended to give generally lower probabilities, yielding better calibration, but did not give judgments that more accurately discriminated between effective and ineffective treatments. In the covariance decomposition, this difference is most apparent in the bias component: The actors systematically overestimated the effectiveness of their treatments to a greater extent than did the observers. The lack of any difference in resolution (or slope in the covariance decomposition) suggests that neither group was better than the other in terms of using the (limited) cues available for judgment in our task (see Yates, 1994).

Differences in the scatter component of the covariance decomposition are also suggestive: Scatter was generally lower when subjects assessed their own treatment recommendations than when they assessed the treatment recommendations of their experimental partner. One possible explanation is that subjects in the latter case faced not only the uncertainty of the patient's responsiveness to treatment, but also uncertainty regarding the reasons the experimental partner recommended the treatment in question. The added uncertainty might have introduced greater variance (unrelated to treatment effectiveness) into their confidence assessments. The influence on overconfidence of social psychological factors involving observers' attributions of actor behavior, and vice versa, is a promising avenue for future investigation.

REFERENCES

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, **78,** 1–3.

Crutchfield, J. P., Farmer, J. D. & Huberman, B. A. (1982). Fluctuations and simple chaotic dynamics. *Physics Reports*, **92,** 45–82.

Diehl, E., & Sterman, J. D. (1995). Effects of feedback complexity on dynamic decision making. *Organizational Behavior and Human Decision Processes*, **62,** 198–215.

Fischhoff, B., & Bar-Hillel, M. (1984). Focusing techniques: A shortcut to improving probability judgments? *Organizational Behavior and Human Performance*, **34,** 175–194.

Fischhoff, B., Slovic, P. & Lichtenstein, S. (1979). Subjective sensitivity analysis. *Organizational Behavior and Human Performance*, **23,** 339–359.

Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology*, **24,** 411–435.

Harvey, N. (1990a). Judgmental control of the behavior of a dynamical system. In K. J. Gilhooly, M. T. G. Keane, R. H. Logie, & G. Erdos (Eds.), *Lines of thinking: Reflections on the psychology of thought*, (pp. 337–352). New York: Wiley.

Harvey, N. (1990b). Effects of difficulty on judgmental probability forecasting of control response efficacy. *Journal of Forecasting*, **9,** 373–387.

Juslin, P., Olsson, H., & Björkman, M. (in press). Brunswikian and Thurstonian origins of bias in probability assessment: On the interpretation of stochastic components of judgment. *Journal of Behavioral Decision Making*.

Keren, G. (1991). Calibration and probability judgments: Conceptual and methodological issues. *Acta Psychologica*, **77,** 217–273.

Koehler, D. J. (1994). Hypothesis generation and confidence in judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **20,** 461–469.

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

Kleinmuntz, D. N. & Thomas, J. B. (1987). The value of action and inference in dynamic decision making. *Organizational Behavior and Human Decision Processes*, **39,** 341–364.

Langer, E. J. (1975). The illusion of control. *Journal of Personality and Social Psychology*, **32,** 311–328.

Lichtenstein, S. & Fischhoff, B. (1980). Training for calibration. *Organizational Behavior and Human Performance*, **28,** 149–171.

Lichtenstein, S., Fischhoff, B. & Phillips, D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases*, (pp. 306–334). Cambridge: Cambridge University Press.

May, R. M. (1986). Simple mathematical models with very complicated dynamics. *Nature*, **261,** 459–467.

Moray, N., Lootstein, P. & Pajak, J. (1986). The acquisition of process control skills. *IEEE Transactions: Systems, Man and Cybernetics, SMC-16*, 497–504.

Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology*, **12,** 595–600.

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

Sterman, J. D. (1989). Misperceptions of feedback in dynamic decision making. *Organizational Behavior and Human Decision Processes*, **42,** 301–335.

Tyebjee, T. T. (1987). Behavioral biases in new product forecasting. *International Journal of Forecasting*, **3,** 393–404.

Received: November 13, 1996

Vallone, R. P., Griffin, D. W., Lin, S., & Ross, L. (1990). Overconfident prediction of future actions and outcomes by self and others. *Journal of Personality and Social Psychology*, **58,** 582–592.

Wright, G. & Ayton, P. (1989). Judgmental probability forecasting for personal and impersonal events. *International Journal of Forecasting*, **5,** 117–125.

Yates, J. F. (1982). External correspondence: Decompositions of the mean probability score. *Organizational Behavior and Human Performance*, **30,** 132–156.

Yates, J. F. (1988). Analyzing the accuracy of probability judgments for multiple events: An extension of the covariance decomposition. *Organizational Behavior and Human Decision Processes*, **41,** 281–299.

Yates, J. F. (1990). *Judgment and decision making*. Englewood Cliffs, NJ: Prentice Hall.

Yates, J. F. (1994). Subjective probability accuracy analysis. In G. Wright & P. Ayton (Eds.), *Subjective probability* (pp. 381–410). New York: Wiley.