Journal of Behavioral Decision Making, Vol. 10, 221–242 (1997)

Confidence Judgments by Actors and Observers

DEREK J. KOEHLER AND NIGEL HARVEY
University College London, UK

ABSTRACT

We report three experiments comparing confidence judgments made by actors and by observers. In Experiment 1, actors generated qualitative answers (countries of the world) in a country-identification task; in Experiment 2, actors generated quantitative answers (years) in a historical event-dating task. Both actors and observers indicated their confidence in the actors' answers. Actors were significantly less confident in their answers than were observers in the first experiment. This effect was substantially reduced in the second experiment, whether confidence was measured by judged probability or by credible interval width. Experiment 3 used a control task in which actors attempted to bring an outcome variable into a desired range. In contrast to the first two experiments, actors in the control task were more confident than observers. Because subjects were generally overconfident in all three experiments, the present results demonstrate that the use of observers can reduce or exacerbate overconfidence depending on the kind of task and the nature of the event or possibility under evaluation. © 1997 by John Wiley & Sons, Ltd.

Journal of Behavioral Decision Making, 10, 221–242 (1997)

No. of Figures: 3 No. of Tables: 5 No. of References: 22

KEY WORDS confidence; subjective probability; hypothesis generation; calibration

When decisions are made in organizational settings, they are typically evaluated by a separate person or group of people in the institution before being implemented. In addition to creating a sense of responsibility or accountability on the part of the decision maker (e.g. Tetlock and Boettger, 1994; Tetlock *et al.*, 1989), this procedure is often justified by an implicit assumption that a person who was not directly involved in making the decision can offer a more impartial and therefore more accurate assessment of its quality. In particular, this 'observer' may be better able to assess the likelihood that the proposed course of action will be successful than is the 'actor' who formulated the course of action. The goal of the present research was to investigate actor–observer differences in judging the probability of possible outcomes.

The two previous studies that are most relevant to this topic have produced apparently conflicting results. Harvey and Ayton (1990) examined actor–observer differences in a simulated clinical decision-making task developed by Harvey (1990), in which subjects played the role either of a psychiatrist making treatment recommendations or of a nurse observing the doctor's actions. After each treatment

CCC 0894–3257/97/030221–22$17.50

Accepted 14 October 1996

was prescribed by the psychiatrist, both subjects estimated the probability of the treatment being effective. The results showed that the observers (nurses) gave less overconfident and better calibrated probability judgments than did the actors (psychiatrists). Results of this study thus suggest that impartial observers may indeed be able to give more accurate confidence judgments than those given by actors directly involved in making the decision.

Two possible causes for the actor–observer differences were proposed by Harvey and Ayton (1990). First, the observers may have been making their own decisions covertly. In cases where their decision matched that of the actor, similar probability judgments would be given, but where the two disagreed, the observers would presumably give lower probabilities. Second, because the actors had to make both a decision and a probability judgment, in contrast to observers who had only to give a probability judgment, the observers may have made better judgments simply because they had greater processing resources available.

A study by Koehler (1994), however, found the opposite pattern of results. Subjects were asked to generate qualitative (i.e. categorical) hypotheses from an open-ended set, for example, a film they believed was most likely to win the Best Picture Oscar. Actors generated a hypothesis of their own, while observers evaluated a hypothesis previously generated by an actor. The results showed that actors actually gave lower confidence judgments than did observers, although the same set of hypotheses was being judged in both conditions. Results using a large number of 'fill-in-the-blank' general knowledge questions showed that actors' probability judgments were not only less overconfident than those of the observers but that they were also better calibrated and had greater resolution (i.e. more clearly distinguished correct from incorrect answers).

These results were interpreted as follows. Assuming that the actor generated several hypotheses before settling on his or her preferred answer (focal hypothesis), these alternatives would still be salient when the probability judgment is given, leading to lower confidence in the focal hypothesis relative to that of the observers. Consideration of alternative hypotheses would also be expected to improve calibration and resolution. Two additional findings support this interpretation. First, when all the alternatives were known to both actors and observers, then the effect was reversed. Actors choosing an alternative from a closed set were more confident in their choice than were observers who evaluated the alternative selected by the actor. Second, when a distractor task was inserted between the time the actor generated a hypothesis and evaluated its probability, the original actor–observer difference was also eliminated. The delay increased the actors' probability judgments so that they roughly equaled those of the observers.

These results are consistent with a descriptive theory of probability judgment, called support theory, developed by Tversky and Koehler (1994). In support theory, probabilities are assigned to descriptions of events (called *hypotheses*) rather than to the events themselves, such that two different descriptions of the same event can receive different probability judgments. Judged probability is assumed to depend on the support that the focal hypothesis and its alternative receive from the available evidence. Support theory emphasizes the individual's representation of the events under consideration, and assumes that a hypothesis receives greater support when it is represented explicitly rather than implicitly as a subset of a more inclusive event. Such an assumption is consistent with research showing that an event is judged as more probable when it is specified explicitly than when it is included implicitly in a residual category (Gettys *et al.*, 1986; Fischhoff *et al.*, 1978; Mehle *et al.*, 1981). In Koehler's (1994) experiments, the observer is assumed to represent the alternative to the focal hypothesis as a single entity, namely as the negation of the focal hypothesis, while the actor is assumed to have 'unpacked' it into a disjunction of the specific alternatives that came to mind during hypothesis generation. As a consequence, the judged probability of the focal hypothesis is lower for actors than for observers.

One of the main purposes of the present research was to investigate the causes of the conflicting results obtained by Harvey and Ayton (1990) and Koehler (1994). In so doing, the research may help to

identify the factors determining when observers are likely to make better or worse probabilistic assessments of decision outcomes than those made by the actors involved in making the decision. The results and interpretations of the previous studies suggest that two opposing differences between actors and observers are at work:

(1) *Discrepant belief*. Actors and observers have different knowledge bases. The major source of this difference is simple person-to-person variability in experience and education. If the actor selects the hypothesis for which he or she has the most evidential support available, then we would expect the observer to have somewhat less evidential support on average due to this variability. Thus, by regression alone, we would expect generally lower probability judgments from the observers than from the actors. Because probability judgments are generally too high for tasks offering at least a moderate degree of difficulty (e.g. Lichtenstein *et al*., 1982), observers' judgments consequently should be less overconfident and better calibrated than those of the actors. No differences would be expected for resolution, however, because there is no reason to believe that the evidence available to observers is any better (i.e. more diagnostic) on average than that available to actors.
(2) *Differential unpacking*. Actors and observers have different representations of the relevant events. The major source of this difference, according to the interpretation offered above, is the actors' tendency to unpack more alternatives from the residual category as a result of making the judgment or decision. This should lead to generally lower probability judgments for actors than for observers. Unpacking of alternatives should not only lower the average probability judgment for the actors, thereby leading to less overconfidence and better calibration, but would also be expected to increase resolution. That is, because the most likely alternatives to the focal hypothesis are unpacked, the resulting probability judgment should be better able to discriminate correct from incorrect judgments, since many of the incorrect judgments would be incorrect precisely because one of the unpacked alternatives turned out to be correct instead.

From this view, confidence was greater for observers than for actors in Koehler (1994) because the effect of differential unpacking dominated that of discrepant belief, while the opposite held in Harvey and Ayton (1990), where differences in representation were less likely, perhaps because the alternatives were fairly well-specified and the observers may have been making covert decisions. The distinction between discrepant belief and differential unpacking not only accounts for the basic discrepancy between the results of the two studies, but is also consistent with (1) the reversal of the actor–observer discrepancy in Koehler (1994) when a fixed set of alternatives is specified for both actors and observers, a manipulation that should eliminate differences due to differential unpacking and allow discrepant belief to dominate; (2) the significant increase in confidence observed by Koehler (1994) when actors completed a distractor task after generating their hypotheses but before assessing their probability, a manipulation that should eliminate differences due to discrepant belief since the same person now plays the role of actor and observer, leaving differential unpacking to dominate; and (3) the significant difference in resolution between actors and observers in Koehler (1994) where differential unpacking dominated but not in Harvey and Ayton (1990) where discrepant belief dominated.

In the present experiments we investigate four manipulations intended to reduce the influence of differential unpacking, that is, differences in problem representation. These manipulations would be expected to increase the impact of discrepant belief and hence the likelihood that actors would be more confident than observers.

(1) In addition to the actor and observer conditions, a third, modified observer condition is also used in all three experiments in which the observer generates a possible alternative to the actor's hypothesis before assessing the likelihood that the actor's hypothesis is correct. This manipulation was expected to reduce differential unpacking by encouraging observer subjects to 'unpack' alternatives to the actor's guess (cf. Dube-Rioux and Russo, 1988; Mehle *et al*., 1981).

(2) In the first two experiments, the standard actor condition was compared with an actor-delay condition in which subjects first provided guesses for the problems and only then went back and gave a probability judgment for each guess (cf. Koehler, 1994, Exps 5 and 6). The assumption was that this delay would reduce the salience of the alternatives unpacked during hypothesis generation, and thereby attenuate the impact of differential unpacking on confidence.
(3) Actors in the first experiment generated qualitative hypotheses as answers, as in the Koehler (1994) study. In the second experiment, in contrast, actors generated quantitative hypotheses as in the Harvey and Ayton (1990) study. Such a manipulation would also be expected to reduce differential unpacking to the extent that it reduces memory limitations (cf. Tversky and Koehler, 1994) by making it easier to unpack alternative hypotheses. That is, it is possible that the use of quantitative hypotheses leads to more spontaneous unpacking of the residual category by observers, hence reducing actor–observer differences due to differential unpacking.
(4) The first two experiments both use general knowledge problems (country identification and historical event dating, respectively), but in the third experiment the Harvey and Ayton (1990) control task is used instead. It is possible that difference of opinion or belief will play a larger role in the case of deciding on a course of action than it does in the case of general knowledge problems; for instance, the fact that there is not necessarily a single correct answer — as there is with the general knowledge items — might exacerbate the effect of discrepant belief.

EXPERIMENT 1

Method

Subjects
Subjects were 72 students, prospective students, and staff from University College London. Prospective students participated as part of a laboratory demonstration; the remaining subjects were paid for their participation. One subject failed to answer a large proportion of the problems; data from this subject and that of the person with whom this subject was paired were dropped from subsequent analysis.

Stimuli
Subjects were presented with 60 countries for identification. For each country they were provided with a picture of the country's boundary shape and four further cues (area, population, literacy rate, and hemisphere) to its identity. The picture of each country's shape did not show any adjacent countries or give any other cues to its relative location. The area cue was given relative to the size of the United Kingdom (e.g. '3.2 × UK' indicated that the country in question has an area 3.2 times that of the United Kingdom). Population was rounded to the nearest million if the population of the country in question is greater than 3 million, to the nearest 100,000 if its population is less than 3 million but more than 1 million, and to the nearest 10,000 if its population is less than 1 million. The literacy rate, which indicated the percentage of people who can read and write in the country, was rounded to the nearest 1%. The hemisphere cue indicated whether the country is located in the northern or southern hemisphere; countries located on the equator were classified on the basis of whether the majority of their area is above or below the equator. Cue information was obtained from the 1994 edition of the *Hutchinson Pocket Factfinder*.

Design
Subjects were assigned to one of five conditions. For each country, subjects in the actor condition ($n = 14$) generated a guess and then assessed the probability that this guess was correct. Subjects in the

observer and modified observer conditions ($n = 14$ each) were yoked with subjects in the actor condition, yielding 14 yoked triplets. Subjects in these conditions assessed the probability that the actor's guess was correct; before doing so, subjects in the modified observer condition first wrote down an alternative answer that might be true instead. Subjects in the actor-delay condition ($n = 14$) went through the questionnaire and wrote guesses for all 60 countries and only then went back through and assigned a probability to each guess. Subjects in the observer-delay condition ($n = 14$) assessed the probability of the guesses provided by subjects in the actor-delay condition with whom they were paired. The 60 country problems were presented in the same fixed order for all subjects.

Procedure

Subjects in all conditions were given identical instructions regarding interpretation of the cues to country identity, in which the meaning of each of the four cues was defined. The instructions warned that the pictures of the country's boundaries were not drawn to scale, and that the area cue (rather than the relative size of the pictures) should be used to obtain an accurate sense of each country's size. Subjects were told that some of the countries would be quite easy to identify while others would be very difficult. They were also informed that correct spelling of the country's name was not important. Subjects in the actor conditions were instructed to write down an answer for every problem, even if it was only a guess; subjects in the observer conditions were informed that this instruction had been given to the actors.

The 60 country problems were presented in a questionnaire packet with two problems per page. Actor subjects wrote their guesses on a blank line provided below each problem. Their original, completed problem forms were later presented to the yoked subjects in the observer conditions, first to subjects in the standard observer condition and then to subjects in the modified observer condition. These subjects were told that they had been paired with a randomly selected subject from an earlier experiment, and would be presented with this previous subject's answers for evaluation. Subjects in the modified observer condition were instructed to write down an alternative answer that might be correct instead below the actor's guess on the original problem form. They were encouraged to write an alternative even if it was only a guess, and even if they believed that the actor's original guess was actually correct.

All subjects made their judgments of the probability that the actor's guesses were correct on a separate response form. For each problem, identified by problem number, they were asked to circle their probability judgment on a scale running from 0% to 100% in intervals of 10%. Subjects in the actor-delay condition first wrote down guesses for all 60 problems and only then were given the probability response form and instructions. All other subjects gave a probability judgment after each problem. In all conditions the following instructions were given regarding the probability judgment task:

> A rating of 100% means you are certain that the answer is correct, and a rating of 0% means you are certain the answer is incorrect. Intermediate ratings indicate intermediate degrees of certainty about the answer's correctness. For example, a rating of 50% means the answer is as likely to be correct as it is to be incorrect; in other words, the probability is the same as tossing a coin and having it come up heads.

Subjects in the observer conditions were not presented with the probability judgments provided by the original actor subjects.

Results and discussion

On average, across subjects and problems, actors in the immediate assessment task answered 32%, or 19 of the 60 problems correctly. The most accurate subject answered 40 of the 60 problems correctly;

Exhibit 1. Results from Experiment 1

Measure	Condition			Omnibus F	Contrast $F(1, 26)$
	Actor	Observer	Modified observer		
Immediate assessment					
Confidence	38%	50%	52%	6.46[a]	12.87[a]
Accuracy	32%	(32%)	(32%)	—	—
Brier	0.115	0.208	0.210	9.58[a]	19.19[a]
Calibration	0.051	0.109	0.105	4.71[b]	9.41[a]
Resolution	0.124	0.089	0.083	5.06[b]	10.06[a]
Delayed assessment					
Confidence	37%	56%	—	21.59[a]	—
Accuracy	29%	(29%)	—	—	—
Brier	0.109	0.242	—	24.37[a]	—
Calibration	0.056	0.158	—	15.65[a]	—
Resolution	0.114	0.084	—	5.53[b]	—

Note: Degrees of freedom for the omnibus test reported above are 2 and 26 for the immediate assessment condition, and 1 and 13 for the delayed assessment condition. The contrast compares the actor condition to the two observer conditions. Accuracy values are in parentheses for observers because this variable was completely determined by actors' responses.
[a] $p < 0.01$.
[b] $p < 0.05$.

the least accurate subject correctly answered only 8 of the 60 problems. Accuracy in the delayed assessment task ($M = 29\%$) was similar to that found in the immediate assessment task, indicating that the elicitation of the confidence judgment had no effect on subjects' accuracy in the country identification task.

The main results are listed in Exhibit 1. As indicated by the table, the actor–observer manipulation had a significant effect on confidence in the actor's best guess. In the immediate assessment task, subjects in the two observer conditions expressed significantly greater confidence in the actors' best guesses than did the actors themselves. In the delayed assessment task, observers were also significantly more confident than the actors with whom they had been paired. This result, then, replicates Koehler's (1994) finding of greater confidence for observers than for actors. In contrast to expectations, however, there was no difference in confidence between the standard and modified observer groups in the immediate assessment task: Observers asked to generate an alternative answer before assessing the actor's guess expressed no less confidence than observers who did not generate an alternative, $F(1, 26) < 1$.

A second unexpected finding was that actors in the delayed assessment task did not differ in terms of their confidence from actors in the immediate assessment condition, $F(1, 26) < 1$. This result is inconsistent with Koehler's (1994, Exps 5 and 6) findings. This might have been due to one of two major differences between the present experiment and those of Koehler (1994). First, a much larger number of problems was used in the present context, all of which were drawn from a single domain (i.e. countries of the world). This might have led in some way to a heightened salience of alternative hypotheses that was not eliminated by the delay. Second, in the present experiment, subsequent problems of the same sort served as the distractor task, while in Koehler (1994) an entirely unrelated task was used as the distractor. Perhaps an unrelated task is needed to fully eliminate salient alternatives from memory. In summary, the delay manipulation might have been ineffective in this context because the distractor task involved generating further hypotheses from the same set, namely countries of the world.

Comparison of mean accuracy to mean confidence indicates that subjects in this experiment were generally overconfident that the actors' best guesses were correct. While actor subjects were, on average, only modestly overconfident, subjects in the two observer conditions were highly overconfident: Their confidence judgments imply that somewhat more than half of the answers should be correct, but in fact less than one in three was correct.

For a more detailed analysis of the relationship between confidence and accuracy, Brier scores (e.g. see Yates, 1990) were computed and decomposed separately for each subject. Exhibit 1 indicates that actors exhibited better overall correspondence (as measured by the Brier score) between confidence and accuracy than did observers (or modified observers). This difference held for both the calibration and resolution components of the Brier score. (By definition, the third component of the Brier score, outcome variance, is the same for the three conditions because the same events were assessed by all three subjects in any triplet.) The actors' confidence judgments, then, both separated correct from incorrect answers more clearly and corresponded more closely to the actual accuracy attained at each confidence level than did the confidence judgments of the observers and modified observers.

Exhibit 2 displays calibration curves for actors and observers (collapsed over the immediate and delayed assessment tasks, and including modified observers). In this and subsequent calibration curves presented in this article, the mean accuracy for all problems assigned a given confidence level is computed separately for each subject and then averaged over subjects within a given experimental condition (i.e. each subject who used a given confidence category contributes equally to that point on the curve). Perfect calibration is indicated by a curve falling along the diagonal. The superior calibration of actors is implied by the fact that their calibration curve is generally closer to the diagonal than that of the observers. The relative flatness of the observers' calibration curve for confidence judgments between 0% and 70% indicates that distinctions in confidence made in this range were only weakly related to accuracy.

EXPERIMENT 2

The second experiment also examined actor–observer differences in the domain of general knowledge problems, but in this case focused on quantitative rather than qualitative hypotheses. If the major source of the apparently contradictory findings of Koehler (1994) and Harvey and Ayton (1990) is the use of qualitative hypotheses in the former and quantitative hypotheses in the latter, then the actor–observer difference found in Experiment 1 should be reversed in this second experiment. If, however, the discrepancy is attributable to the use of a control task by Harvey and Ayton (1990), in contrast to the general knowledge and prediction tasks used by Koehler (1994), then the results should resemble those of the first experiment. Experiment 2 also compares delayed and immediate assessments of confidence — as in the first experiment — but manipulates this factor within rather than between subjects.

Method

Subjects

Subjects were 132 undergraduate and prospective undergraduate psychology majors at University College London, who participated as part of a laboratory demonstration. Data from two additional subjects (and the subjects with whom they had been paired) were dropped as these subjects failed to complete the task as instructed.

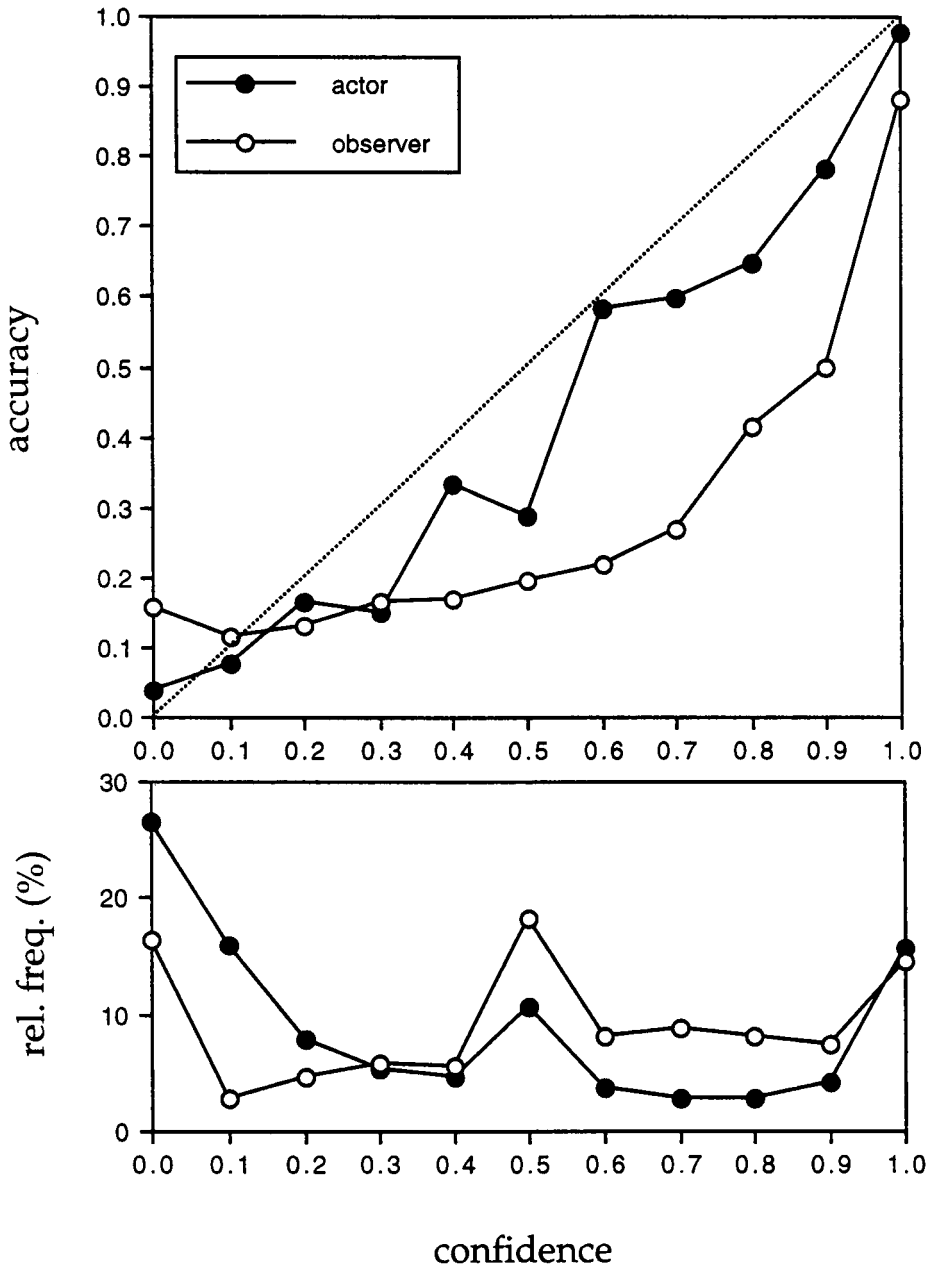

Exhibit 2. Calibration curves and relative frequency distributions of confidence for actor and observer conditions of Experiment 1

Stimuli

Subjects were presented with 100 well-known historical events with the task of estimating the year in which each event occurred. Most of the events were political or cultural in nature, and were selected from the *Hutchinson Pocket Factfinder* with a British audience in mind. Examples include 'Colonization of Australia begins' (1788), 'Empire State Building in New York City completed' (1931),

'Domesday Book of England completed' (1086), and 'Monet paints ''Impression, Sunrise'' ' (1872). All events took place in years AD. The earliest event presented occurred in 43 AD ('Romans invade Britain'); the most recent occurred in 1979 ('Nuclear reactor accident at Three Mile Island, USA'). The distribution of events was skewed toward relatively recent events. Approximately 60% of the events occurred after 1800, and another 15% between 1600 and 1800.

Design

As in Experiment 1, yoked triplets of actors, observers, and modified observers ($N = 44$ triplets) were used. In this experiment, however, the delay factor was manipulated within rather than between subjects as in the previous experiment. The 100 events were presented to all subjects in the same fixed order. Among the actor subjects, half gave immediate confidence judgments for the first 50 events and delayed confidence judgments for the second 50; the other half gave delayed judgments for the first 50 and immediate judgments for the second 50. As in the previous experiment, the delay was introduced by asking subjects to give a guess for all the problems in the set and only then asking them to go back and give a probability for each answer. Subjects in the observer and modified observer conditions all gave immediate confidence judgments.

The way in which confidence was elicited was varied between subjects. For half of the yoked triplets of subjects ($n = 22$ triplets), confidence was assessed via judged probability as in the previous experiment. For the other half of the yoked triplets ($n = 22$ triplets), confidence was elicited by asking subjects to set 90% credible intervals.

Procedure

The task and instructions were presented on an IBM PC-compatible computer. All subjects were given identical information regarding the 100 events to be assessed. They were told that all the events had taken place in years AD such that the correct answer for each event was between 0 and 1995 (the year in which the experiment was conducted). They were warned that they might find the task quite difficult in the sense that they might have only a very rough idea of when many of the events occurred.

Subjects in the actor condition were required to type in a number between 0 and 1995 for every event. As in the previous experiment, observer subjects were told that they had been paired with a previous subject selected at random. Subjects in the modified observer condition were requested to type in their own guess as to when each event occurred after seeing the actor's guess. In contrast to the previous experiment, they were free to type in the same answer as that provided by the previous subject if they so desired. (We felt it less likely in this experiment than in the previous one that subjects in the modified observer condition would be tempted to simply repeat the actor's guess whenever a difficult problem was encountered.)

Subjects for whom confidence was measured by judged probability were asked to assess the probability that the correct year was included in an interval created by the computer around the actor's best guess. This interval was bounded by the actor's best guess plus or minus an integer error term E, where

$$E = (2000 - \text{best guess})/8$$

The value of E was rounded to the nearest multiple of 5 for $E > 20$. Whenever the resulting value of E implied a lower bound less than zero, the lower bound was set to zero; likewise, resulting upper bounds greater than 1995 were set to 1995. The equation yields intervals that become wider the longer ago the event is believed to have taken place. This process of constructing intervals for judgment was used to avoid potential floor or ceiling effects that might result if a fixed interval width was used for all events.

Because the interval was constructed on the basis of the actor's best guess, all three subjects in each yoked triplet assessed the same interval for each event. Note that this interval might or might not include the alternative guess provided by subjects in the modified observer condition. As in the previous experiment, subjects gave their probability judgments using a scale running from 0% to 100% in intervals of 10%. This scale appeared on the computer screen with a box around the 50% mark; subjects moved this box to the left or right using the arrow keys on the keyboard and pressed enter when it was on their selected probability judgment. Instructions regarding probability judgment were the same as in the previous experiment.

Subjects for whom confidence was measured via credible intervals were asked to set low and high bounds around each best guess. They were instructed to type in a low bound such that there was only a 5% chance that the correct year was lower than the low bound, and a high bound such that there was only a 5% chance that the correct year was greater than the high bound. Taken together, they were told, the bounds should form an interval which they were 90% certain included the correct year. The instructions noted that narrower intervals indicated greater certainty that the actor's best guess was correct. Subjects in all three conditions were constrained to include the actor's best guess in the credible interval. That is, attempts to enter a low bound higher than the best guess or a high bound lower than the best guess were not allowed; in these cases the computer beeped and prompted the subject to include the best guess within the bounds. Subjects in the modified observer condition were further constrained to include their own guess — as well as that of the actor with whom they had been paired — within their credible interval.

Results and discussion

The manipulation of immediate versus delayed confidence judgments had no significant effect on the actors' confidence judgments in either the probability judgment or the credible interval condition, $F(1, 21) < 1$ in both cases. To assess the possibility of any carryover effects, the analysis was also conducted for the first half of the task only, but again no difference between the two conditions was found, $t(20) = 0.11$ and 1.12 for the probability judgment and credible interval tasks, respectively, n.s. Because the magnitude of the effect was trivial in both conditions, this variable is ignored in the analyses that follow. Further studies may be needed to identify the source of the discrepancy between the present results and those of Koehler (1994).

Results from the probability judgments, which are listed in Exhibit 3, are considered first. In contrast to the results of the previous experiment, there was no significant difference in confidence between

Exhibit 3. Probability judgment results from Experiment 2

Measure	Condition			Omnibus $F(2, 42)$	Contrast $F(1, 42)$
	Actor	Observer	Modified observer		
Confidence	43%	47%	48%	1.11	2.17
Accuracy	23%	(23%)	(23%)	—	—
Brier decomposition					
Brier	0.227	0.270	0.274	3.16[a]	6.34[b]
Calibration	0.087	0.125	0.130	2.81[a]	5.49[b]
Resolution	0.026	0.022	0.022	0.79	1.65

Note: Contrast compares the actor condition to the two observer conditions. Accuracy values are in parentheses for observers because this variable was completely determined by actors' responses.
[a] $p < 0.10$.
[b] $p < 0.05$.

actors, observers, and modified observers. While the difference between actors and observers is in the same direction as in Experiment 1, the magnitude of this difference (0.04) is quite small when compared to that of the first experiment (0.12). It is possible, then, that the use of quantitative hypotheses may have contributed to the elimination of the actor–observer difference found using qualitative hypotheses in the previous experiment.

Not only did actors not differ from observers, but — once again — the modified observer group showed no effect of generating an alternative hypothesis relative to the standard observers, in contrast to our initial expectations. This was true even though, on average, the alternative guess provided by the modified observers generally improved upon the original guess of the actor: The mean absolute deviation of the guess from the correct year was 155 years for the actor's guess but only 126 for the alternative produced by the modified observers, $t(99) = 5.46$, $p < 0.01$. To assess the effect of the actor's guess on the modified observer's subsequent guess, the correlation between the two guesses was computed across the yoked subject pairs separately for each of the 100 events. That is, for a given historical event, the correlation between the actor's guess and the modified observer's guess was computed across the 22 subject pairs. Correlations of zero are expected if the actor's estimate has no influence; positive correlations indicate that the modified observers' estimates were influenced in the direction of the actors' initial estimates. The average correlation for the 100 items was 0.48, with 62 of the 100 correlations being significantly greater than zero at $p = 0.05$. Modified observers, then, may have used the actor's estimate as an anchor (either intentionally or unintentionally), which was then adjusted whenever they thought they could improve on the initial guess.

On average, across subjects and events in the judged probability condition, the correct year fell within the interval set by the computer around the actor's best guess with a relative frequency of 23%. Mean accuracy (by this measure), then, was less than the mean confidence expressed by all three groups, indicating substantial overconfidence that the interval included the correct year, $p < 0.001$ by paired t-test for all three groups. Individual accuracy varied considerably from subject to subject within the actor condition: The best-performing subject achieved 50% accuracy and the worst achieved only 6%.

Brier scores were computed and decomposed separately for each subject. The mean values are shown in Exhibit 3. This analysis reveals significantly better correspondence between confidence and accuracy for actors than for the two observer groups. Decomposition of the Brier score into calibration and resolution components indicates that the difference is attributable to somewhat better calibration (and not better resolution) for actors than for observers. This calibration difference is shown in Exhibit 4, which displays the calibration curves for the actor and observer conditions (observer and modified observers have again been combined). Thus, although there was no direct effect of the actor–observer manipulation on mean confidence, this manipulation did affect the relationship between confidence and accuracy.

We now turn to analysis of the credible interval data. Here, the width of the interval set by the subject is taken as a measure of that subject's confidence in the actor's best guess. The mean results are shown in Exhibit 5. (The width measure was first transformed to a logarithmic scale before averaging to reduce the disproportionate effect of very high values.) As indicated in the table, the actor–observer manipulation had a small but nonsignificant effect on confidence as measured by interval width. As in the judged probability condition, actors were somewhat less confident in their best guesses (i.e. set wider intervals) than the observers or modified observers. Thus, although the effect did not reach statistical significance in either case, the direction of the actor–observer difference was the same whether confidence was measured by judged probability or by credible interval width.

Subjects in the modified observer condition appeared to be somewhat less confident than subjects in the standard observer condition, although this effect was quite small and not statistically significant, $F(1, 42) < 1$. It is worth noting this difference, however, as it is the only indication obtained in the

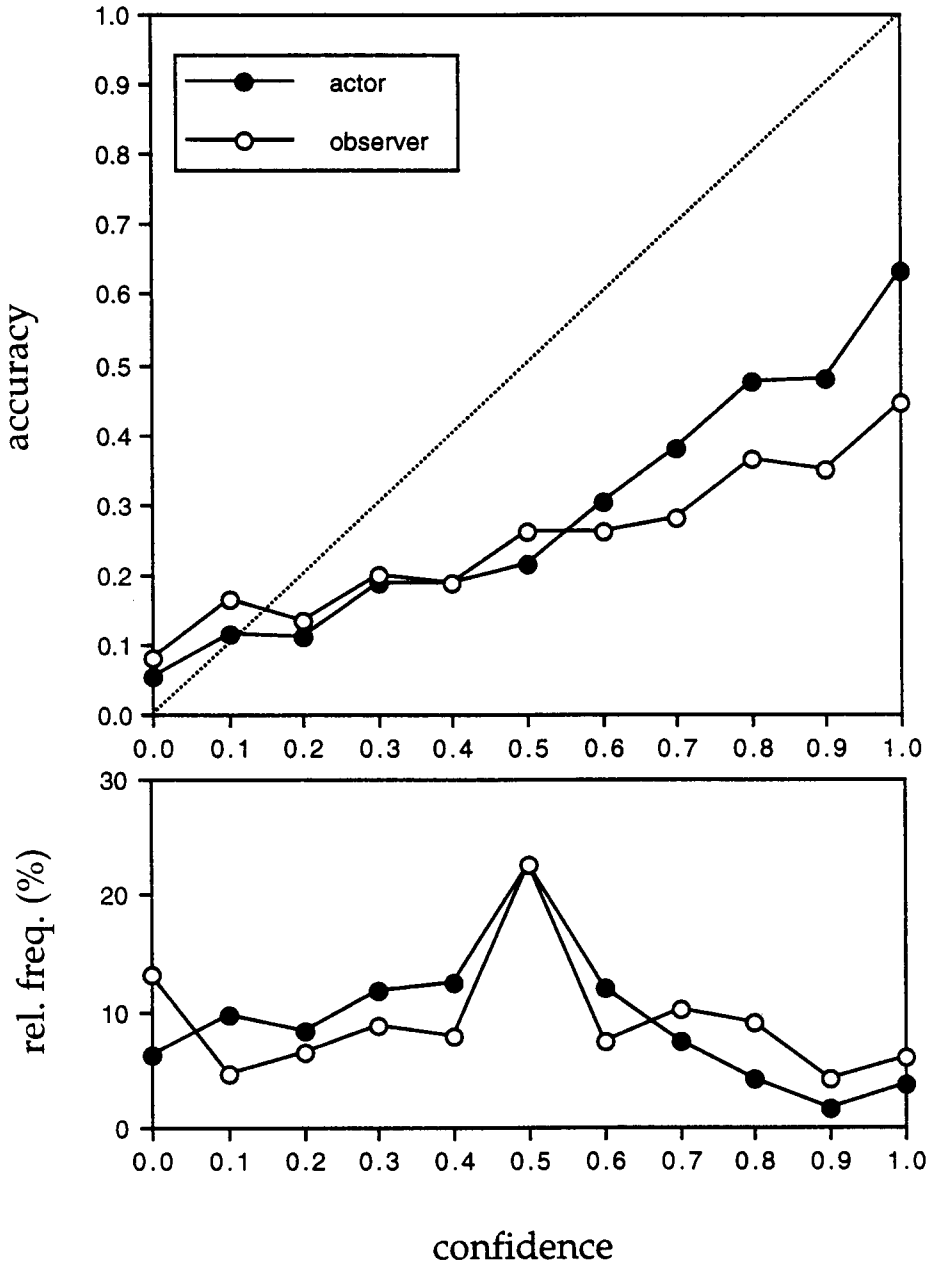

Exhibit 4. Calibration curves and relative frequency distributions of confidence for actor and observer conditions of Experiment 2

present studies of any effect of generating an alternative on confidence in the actor's guess. It may be that what little effect results from generating an alternative hypothesis is found only when confidence is measured using credible intervals rather than judged probability.

Again, as in the judged probability condition, subjects in the modified observer condition gave more accurate alternative answers than those of the original actor subjects, $t(99) = 4.28$, $p < 0.01$. Guesses

Exhibit 5. Credible interval results from Experiment 2

Measure	Condition			Omnibus $F(2, 42)$	Contrast $F(1, 42)$
	Actor	Observer	Modified observer		
Confidence (interval width)	61	47	53	1.22	1.94
Accuracy (hitrate)	37%	38%	36%	0.06	0.01

Note: The contrast compares the actor condition to the two observer conditions.

from the actor subjects had an average absolute deviation from the correct year of 170 years; the corresponding value for the modified observer subjects was 144 years. Analysis of the correlation between actors' and modified observers' guesses (computed over subject pairs separately for each event) again indicated a significant influence of the actors' initial guess on the modified observers' subsequent guess: The average correlation between the two was 0.53; of the 100 correlations, 63 were significantly greater than zero at $p = 0.05$.

Exhibit 5 also lists the relative frequencies with which the intervals set by subjects in the three conditions actually included the correct year. Ideally, this proportion — sometimes referred to as a hitrate — should be 90% in all conditions as subjects had been asked to set 90% credible intervals. As shown in the table, however, subjects in all three conditions failed to achieve this goal. As has been found in much previous research (e.g. Alpert and Raiffa, 1982; Peterson and Pitz, 1988), subjects tended to set intervals that were too narrow, with the result that they in fact included the correct year in fewer than 40% of the judgments. Individual analysis showed that this relative frequency value was below 90% for every subject and for each event included in the experiment. Thus, both the credible interval and the judged probability measures of confidence suggest that subjects in this task were generally overconfident that the actors' best guesses were accurate. The proportions of intervals including the correct year did not differ significantly between actors, observers, and modified observers.

EXPERIMENT 3

All else being equal, and assuming that there exist differences in belief among individuals, the discrepant belief interpretation outlined in the introduction predicts that observers should express less confidence in a hypothesis generated by an actor than that expressed by the actors themselves. The first experiment showed, as did the studies of Koehler (1994), that when generating qualitative hypotheses, actors are actually less confident than observers. The second experiment showed that this difference is essentially eliminated when quantitative hypotheses are used instead, but still failed to find *greater* confidence for actors than for observers as was reported by Harvey and Ayton (1990). The use of quantitative hypotheses, then, can only partly account for the discrepancy in results between Koehler (1994) and Harvey and Ayton (1990).

In Harvey and Ayton (1990), observers sat beside the actors and watched as they made their decisions, at which point both evaluated the probability that the outcome variable would fall in the desired range. As noted by Harvey and Ayton (1990), this procedure allows the possibility that the observers were covertly making their own decision while waiting for the actors to make theirs. To test this possibility, Experiment 3 employed the same task used by Harvey and Ayton (1990), but with an experimental design analogous to that of the two previous experiments. Because observers were presented concurrently with the outcome of the previous trial and the actor's decision regarding the next trial, they did not have an opportunity to process the feedback and form their own opinions about

the next trial before seeing the actor's decision. (Obviously, observers could still make their own covert decisions before assessing that of the actor, but this could be viewed as strategic behavior, which differs somewhat from the interpretation that the delay between feedback and the actor's decision induces observers to make covert decisions that they otherwise would not have made.) The interpretation of the original Harvey and Ayton (1990) study in terms of covert decision making by observers would be supported by a reduction or elimination of the greater confidence for actors observed in their original study. If actors were to remain more confident than observers, however, then we may have to conclude that actor–observer differences depend critically on the type of task under study.

Method

Subjects

Subjects were 105 undergraduate and prospective undergraduate psychology majors at University College London, who participated as part of a laboratory demonstration.

Stimuli

Subjects were presented with eight 'patients', each of whom was treated for 10 consecutive months. Each month the patient's 'happiness index' (which could range from 0 to 50) was presented; the goal was to bring this index within the desired range of 29–31. The patient's behavior was controlled by the following noisy logistic map:

$$B_i = A\ B_{i-1}(B_{i-1} - 1) + e$$

where B_i is the patient output for month i ($0 < B_i < 1$), A is a control parameter, and e is a normally distributed error term with a mean of 0 and a standard deviation of 1. The value of B_i was multiplied by 50 to obtain the happiness index value. The eight patients started with different values of the parameter A, which controls the difficulty of treating the patient (greater difficulty is associated with greater values of A, see Harvey, 1990). Values of A greater than 1.0 but less than 3.0 yield stable output values that increase with A. Values greater than 3.0 but less than 3.57 yield output alternating between two stable values. Values of 3.57 or greater yield unpredictable, chaotic output. The value of the A parameter for the eight patients, along with 12 months' sample output in the absence of any treatment, is presented in Exhibit 6. The initial state for each patient was determined by choosing a random starting value between 0 and 1 for B_i and then cycling the system through 200 iterations, after which the first two months' output was generated for presentation to the subject.

Exhibit 6. Sample happiness index values for 12 consecutive months in the absence of treatment, listed for each of the eight values of the A parameter (i.e. 'patients') used in Experiment 3

A	Sample output (months 1–12)											
2.0	24	24	25	24	24	25	25	25	24	24	25	24
2.2	28	26	27	26	28	28	26	26	27	27	27	26
2.9	33	32	32	32	33	31	33	31	33	31	33	32
3.1	38	27	38	27	38	27	38	27	38	26	38	27
3.3	24	41	21	39	26	41	24	40	25	41	22	41
3.5	19	40	26	43	20	41	23	43	18	39	28	43
3.7	14	38	32	42	22	45	16	42	23	45	15	40
3.9	28	46	13	39	31	45	15	41	27	47	9	29

Subjects could adjust the value of A for a given patient by prescribing lithium, which decreased the value of A, or antidepressant, which increased the value of A. The prescribed dose determined the magnitude of the change: For every milligram of the prescribed drug (which could range from 0 to 30), the value of A was changed by 0.05 in the appropriate direction. For each patient there existed an ideal dose such that, if prescribed consistently, the patient's happiness index would be brought into the desired range (although even then the error term could occasionally drive the output slightly out of range).

Design

Subjects in the actor condition ($n = 35$) were presented with all eight patients, the order of which was determined randomly for each subject. For each patient, actors made a prescription decision and then judged the probability it would be effective; this process was repeated for 10 consecutive 'months' (i.e. trials). All actor subjects gave their probability judgments immediately after making the prescription — no delay condition was used in this experiment. Once again, subjects in the observer and modified observer conditions ($n = 35$ each) were paired randomly with subjects from the actor condition. Each subject in these conditions was presented with the eight patients in the same order as that used for the actor with whom he or she was paired. Subjects in these observer conditions were presented with the same feedback regarding the treatment outcome as that presented to the actors.

Procedure

Subjects in all conditions were given essentially identical instructions regarding the treatment of the patients they would be seeing. They were told that the ideal happiness index score for a patient is 29, 30, or 31, and that patients with indices outside this range would need to be treated with either antidepressant or lithium. Depressed patients, they were told, whose indices were consistently below 29, would need antidepressant. Manic patients, in contrast, whose indices were consistently above 31, would need lithium. Subjects were told that the further the patient's index was from the ideal range, the greater was the dosage required to correct the imbalance. They were told that manic-depressive patients whose indices alternated (either systematically or unpredictably) between depressed and manic would also need lithium. Again, they were told, greater doses would be needed for patients exhibiting more extreme indices. Subjects were told that, for each patient, there was an ideal dose that — if administered consistently — would bring the patient's index into the ideal range. They were warned, however, that such a dose might have to be maintained for more than a single month before it would be effective.

Subjects in the actor condition were presented with the patient's index values for two months preceding treatment. They were then asked to choose whether to prescribe antidepressant, lithium, or no drug. If they chose either drug they were then asked to prescribe a dose between 0 and 30 mg. Subjects in the observer condition were presented with the same data as the actors, along with the treatment prescribed by the actor they were paired with. Subjects in the modified observer condition received similar information, but were also asked to state what they themselves would have prescribed before giving a probability judgment. They entered their own prescription on the screen in boxes adjacent to those presenting the actor's original prescription. If they chose, they were allowed to enter a treatment identical to that of the actor.

All subjects then judged the probability that the actor's prescribed treatment would be effective in bringing the patient's index into the desired range in the next month. Instructions and rating scale were identical to that of the previous experiment. After entering the probability judgment, the subject then

Exhibit 7. Probability judgment results from Experiment 3

Measure	Condition			Omnibus $F(2, 68)$	Contrast $F(1, 68)$
	Actor	Observer	Modified observer		
Confidence	60%	53%	55%	3.48[a]	6.16[a]
Accuracy	27%	(27%)	(27%)	—	—
Brier decomposition					
Brier	0.292	0.278	0.276	0.35	0.69
Calibration	0.158	0.146	0.143	0.26	0.50
Resolution	0.031	0.034	0.033	0.26	0.44

Note: The contrast compares the actor condition to the two observer conditions. Accuracy values are in parentheses for observers because this variable was completely determined by actors' responses.
[a]$p < 0.05$.

was presented with information regarding the patient a month later. This information was presented in a 2-month 'window' format which always displayed the recommended treatment and resulting happiness index from the previous month (i.e. the results of the most recent treatment) and from the month before that. Thus all subjects received the same feedback regarding the treatment outcome. This process continued until the patient had been treated for 10 consecutive months. After the tenth month's confidence judgment was entered, subjects received feedback regarding the outcome along with a message indicating that they had completed the treatment of that patient and would now be seeing a new patient. In this way, subjects made judgments regarding all eight patients.

Results and discussion

The probability judgments were analyzed using a 3 (actor versus observer versus modified observer) by 8 (patient difficulty) by 10 (trial number) repeated measures analysis of variance. The results, listed in Exhibit 7, indicate a significant main effect of the actor–observer manipulation. As in Harvey and Ayton (1990), subjects in the actor condition expressed greater confidence that their prescribed treatment would be effective than did observers or modified observers. Contrasts indicated that the actors differed significantly in terms of confidence from the two observer groups, $F(1, 68) = 6.16$, $p < 0.05$, which in turn did not differ significantly from one another, $F(1, 68) < 1$. The magnitude of the actor–observer difference (0.07), however, is somewhat smaller than the difference (0.11) found in the original Harvey and Ayton (1990) study. This result suggests that while part of the original effect reported by Harvey and Ayton (1990) may be attributable to covert decision making on the part of the observers, results using this control task nevertheless seem to differ substantially from those found using the general knowledge tasks.

On average, across patients and subjects, the actors' prescriptions were successful in bringing the happiness index into the desired range in approximately one out of every four treatments. Thus the probability judgments of all three groups were, on average, higher than the actual relative frequency of successful treatments, indicating overconfidence. By this measure, then, in contrast to the two previous experiments, actors were more overconfident than observers. Exhibit 7 also lists the Brier score results, which were computed and decomposed separately for each subject and then averaged over subjects within a given group. This analysis yielded no significant differences among the three conditions for the Brier score or its components. The lack of a difference in resolution is of particular interest; as suggested by the analysis outlined in the introduction, we find actor–observer differences in resolution only when differential unpacking dominates discrepant belief, that is, only when observers are more

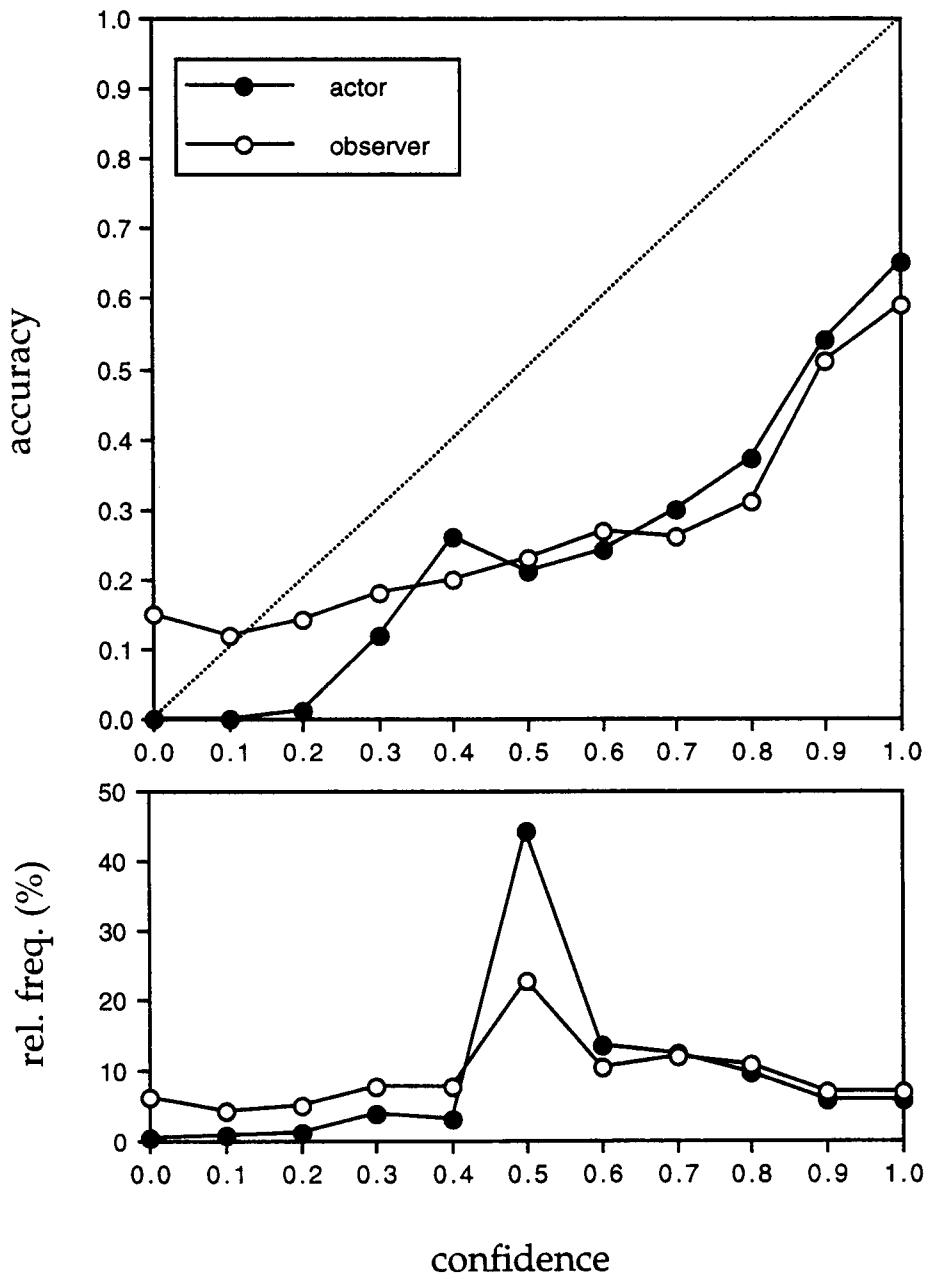

Exhibit 8. Calibration curves and relative frequency distributions of confidence for actor and observer conditions of Experiment 3

confident than actors. Exhibit 8 shows the calibration curves for the actors and observers (again combining data from observers and modified observers). The figure shows generally poor calibration in both conditions, and suggests that actors and observers differed most in their use of the lower end of the probability scale.

GENERAL DISCUSSION

Across the three experiments reported in this article, we observed lower, equal, and higher confidence for actors than for observers evaluating an identical set of hypotheses. The experiments were conducted to explain the apparently discrepant results of Koehler (1994) and Harvey and Ayton (1990) and, in the end, perhaps their major accomplishment was to replicate the discrepancy even after a number of experimental differences were eliminated. Two potentially important differences between the task used in Experiment 3 and those used in Experiments 1 and 2 are that in the former task but not in the latter (1) actors had some control over the value of the outcome variable, and (2) subjects received outcome feedback after every trial. The influence of the latter has been at least partially supported by subsequent research (Harvey *et al.*, 1996). Further research will be needed to disentangle and test these two possibilities. One intriguing observation from the present findings is that observers' mean confidence was quite close to 50% in all three experiments, while actors' mean confidence varied considerably across experiments. From this perspective, the question becomes one of identifying those factors influencing actor but not observer confidence.

The present results also highlight the distinction between qualitative and quantitative hypotheses and how they are evaluated. The greater confidence expressed by observers than by actors in Experiment 1 and in Koehler (1994) when qualitative hypotheses were assessed was all but eliminated when quantitative hypotheses were used instead in Experiment 2. We suggest that this is because the difference between how actors and observers represent the focal hypothesis and its alternative (referred to as an *evaluation frame* by Tversky and Koehler, 1994) is greater for qualitative hypotheses than it is for quantitative hypotheses. The original results of Koehler (1994) were interpreted as suggesting that observers judged the probability of the focal hypothesis relative to its unelaborated complement; actors, in contrast, were assumed to have partially 'unpacked' the complement into its components in the process of choosing the single hypothesis they thought was most likely, yielding lower confidence in the focal hypothesis.

The continuous nature of quantitative hypotheses, in contrast, makes some ways of representing them more natural than others. In Experiment 2, when an interval was generated by the computer around the actor's best guess, it seems natural to represent the judgment as the probability of the value falling within the interval rather than above or below it, regardless of the extent to which various possible best guesses are still on the judge's mind at the time of judgment. This argument implies that the actor isolates or keeps unpacked those qualitative hypotheses that came to mind in the process of hypothesis generation, but does not for quantitative hypotheses, which are instead 'packed' back into a more inclusive residual hypothesis. The use of quantitative hypotheses, consequently, may tend to minimize actor–observer differences in confidence.

Comparisons such as those above must be made across the experiments reported in this article. As a result, our interpretation of the results must be considered speculative as the substantial differences among the tasks used in these experiments hinder any more definitive conclusions. For example, the control task of Experiment 3 differs from the general knowledge tasks of the first two experiments not only in terms of perceived control and feedback, but in many other ways as well. For example, by virtue of the probabilistic nature of the control task, more than one possible treatment decision could yield a successful outcome while by definition there is only one correct answer for a general knowledge question. Furthermore, observers in the general knowledge tasks could assess the hypothesis solely on the basis of their own knowledge, while the control task would seem to require them to try to figure out why the actor has proposed a given course of action. The country-identification task of Experiment l also differs in many ways from the historical event dating task of Experiment 2 in addition to the type of hypothesis — qualitative or quantitative — being assessed. They differ not only in terms of the domain of knowledge being assessed but also in the sense that a single hypothesis (i.e. a country) is

produced and assessed in country identification while a set of hypotheses (i.e. years) is produced and assessed in historical event dating. Social psychological factors may also play a role. Actors may have given lower confidence judgments than observers in the general knowledge tasks because they anticipate a larger share of the blame if a highly confident answer turns out to be wrong; course-of-action decisions, in contrast, may have been perceived as requiring high confidence on the actor's part to justify the decision. While we recognize that such substantial task differences leave open a number of alternative explanations for the differences we observed among the three experiments, we hasten to point out that this set of experiments represents a preliminary attempt to explain the contradictory findings of Harvey and Ayton (1991) and Koehler (1994), who used very different tasks in their original studies. Subsequent research, we hope, will eventually narrow down these task differences to those that are conceptually most important in accounting for variance in actor–observer differences across tasks.

One finding was surprisingly consistent across the three tasks examined in the present research. The modified observer condition, in which subjects generated an alternative before assessing the actor's best guess, did not yield any difference in judged probability from the standard observer condition in any of the three experiments. We conclude with some speculations regarding the failure of this manipulation to reduce confidence (and, consequently, overconfidence) relative to the standard observer condition.

A likely explanation is that observers, even without prompting, attempted to answer each question for themselves, either because the mere presentation of the question elicited a fairly automatic attempt to search for an answer in memory or because subjects intentionally answered the question for themselves as an aid to assessing the previous answer. As a result, standard and modified observers might not be expected to differ systematically. If this is the correct interpretation, then the puzzling issue raised by these results is why observers of either kind were not less confident than the actors in the first two experiments, especially as the second experiment showed that the modified observers were able to provide alternative answers that actually improved upon those of the actors.

The greater confidence for observers than for actors observed in Experiment 1 and by Koehler (1994) has been taken as evidence that, as a result of the hypothesis generation task, actors tend to unpack the alternative or residual category to a greater degree than do observers. Such an interpretation would seem to suggest that, when observers generate an alternative to the actor's guess before giving a confidence judgment, their engagement in the process of hypothesis generation should in effect place them in the same role as the actors. This argument assumes, however, that the hypothesis generation process is essentially identical for actors and observers, that is, that the observers are unaffected by exposure to the actor's guess. The correlational analysis reported in Experiment 2, however, implies that this assumption is incorrect: Having seen the actor's guess, modified observers tended to give alternative answers that were influenced in the direction of the actor's guess. As a result, modified observers tended to generate alternatives that were closer to the guess of the actor with whom they had been paired than would be expected had they not first seen this previous guess.

This point is made clearly by the following analysis of the data from Experiment 2. First, the 'distance' (absolute deviation) was computed between each modified observer's guesses and those of the actor with whom the observer had been paired. This analysis yielded a mean distance (averaged across items and subject pairs) of 82 years for the probability judgment task and 87 years for the credible interval task. To compare this value with what would be expected had the alternative been generated without knowledge of the actor's guess, we computed the average distance between the guesses of each actor and that of all the other actors within each assessment condition. This yielded an average distance between actors' guesses of 167 years for the probability judgment task and 184 years for the credible interval task. The modified observers, then, generated guesses that were, on average, only about half as far away from the actors' guesses as would be expected had they not first seen the actors' guesses, $t(21) = 10.44$ and 13.87 in the probability judgment and credible interval conditions, respectively, both $p < 0.01$.

If the observers fail to recognize the impact of having seen the actor's guess, and assume instead that the alternatives they generate are an 'independent sample' unbiased by the actor's guess, then they would be expected to overestimate the extent to which their own generated alternative supports the hypothesis that the actor's guess is close to the correct value. That is, the influence of the actor's guess may lead the observer to generate an alternative that is closer to, and thus more supportive of, the actor's guess, yielding greater confidence in the actor's guess than would have been the case had the observer not seen it before generating an alternative (for related work, see Kelley and Lindsay, 1993). Under such circumstances, the observer's prompted or spontaneous generation of an alternative guess may not have the predicted effect of reducing confidence in the actor's guess.

Indirect evidence for this interpretation is offered by comparing the results of Experiment 3 with those of Harvey and Ayton (1990). We noted that, in the original Harvey and Ayton (1990) study, the actors and observers worked on each problem in parallel and that, as a result, the observers might have spontaneously made their own covert judgment while waiting for the actors to reach a decision. In Experiment 3, in contrast, observers were presented with the case information and the actor's decision simultaneously, and thus did not have the opportunity while waiting for the actor's decision in which to make a covert, independent judgment. Relative to the original Harvey and Ayton (1990) study, the magnitude of the actor–observer difference was reduced by half. This was true even for the modified observers (who again did not differ from the standard observers), suggesting that the opportunity in the original study for observers to generate an alternative (albeit covert) decision *before* seeing the actor's decision led to a greater reduction in confidence than did the explicit prompt to generate an alternative decision *after* seeing the actor's decision. Such a result is consistent with the interpretation outlined above, but further research is clearly needed. The most obvious test is simply to manipulate whether modified observers generate an alternative before or after seeing the actor's decision or guess.

Implicit in this interpretation is the assumption that observers treat the actors' guesses not merely as target hypotheses for evaluation, but also in a sense as evidence about the likelihood that the hypothesis is correct. That is, if the actor happens to have generated the same hypothesis as that which comes to the observer's mind (or one that is very close in the quantitative hypothesis task), then the observer may be especially confident that the hypothesis is correct. Agreement or disagreement, then, is taken as a cue to likelihood. Indeed, if the observer produces his or her hypothesis independently of the actor, then answers upon which both subjects agree are in fact more likely to be correct than are answers about which they disagree. Under certain circumstances, however, observers' use of the agreement cue may actually lead to poorer correspondence between confidence and accuracy than that achieved by actors, namely if the observers either (1) overcompensate when they agree with the actor or (2) overestimate the extent of their agreement with the actor. The latter possibility may come about if exposure to the actor's hypothesis prevents the observer from generating his or her own hypothesis independently, as discussed above.

This issue has substantial implications for 'debiasing' techniques (see Fischhoff, 1982) used to improve probability judgments. To date, the most effective and widely endorsed prescriptions for improving judgments made under conditions of uncertainty involve prompting the judge to consider more carefully alternatives to the focal hypothesis or course of action (e.g. Hoch, 1985; Koriat *et al.*, 1980; Lord, 1984; Slovic and Fischhoff, 1977). For example, several studies have found that the underestimation of the likelihood of the residual category in the 'fault-tree' task observed by Fischhoff *et al.* (1978) is substantially reduced by encouraging subjects to 'unpack' specific alternatives included in the residual category before giving their judgments (Dube-Rioux and Russo, 1988; Mehle *et al.*, 1981). The present results, however, call into question the efficacy of such a strategy under circumstances in which the alternatives brought to mind are influenced by exposure to the focal hypothesis. In such a case, the prescription to consider alternatives may entail the risk of doing as much harm as good.

AUTHOR NOTES

The research reported in this article was supported by grant R000221383 from the Economic and Social Research Council of the United Kingdom. Address all correspondence to Derek J. Koehler, now at the Department of Psychology, University of Waterloo, Waterloo, Ontario, N2L 3G1, Canada (e-mail: dkoehler@watarts.uwaterloo.ca).

REFERENCES

Alpert, M. and Raiffa, H., 'A progress report on the training of probability assessors', in Kahneman, D., Slovic, P. and Tversky, A. (eds), *Judgment under Uncertainty: Heustics and biases* (pp. 294–305), Cambridge: Cambridge University Press, 1982.

Dube-Rioux, L. and Russo, J. E., 'An availability bias in professional judgment', *Journal of Behavioral Decision Making*, **1** (1988), 223–37.

Fischhoff, B., 'Debiasing', in Kahneman, D., Slovic, P. and Tversky, A. (eds), *Judgment under Uncertainty: Heuristics and biases* (pp. 422–44), Cambridge: Cambridge University Press, 1982.

Fischhoff, B., Slovic, P. and Lichtenstein, S., 'Fault trees: sensitivity of estimated failure probabilities to problem representation', *Journal of Experimental Psychology: Human Perception and Performance*, **4** (1978), 330–44.

Gettys, C. F., Mehle, T. and Fisher, S., 'Plausibility assessments in hypothesis generation', *Organizational Behavior and Human Decision Processes*, **37** (1986), 14–33.

Harvey, N., 'Effects of difficulty on judgemental probability forecasting of control response efficacy', *Journal of Forecasting*, **9** (1990), 373–87.

Harvey, N. and Ayton, P., 'Actor–observer differences in judgmental forecasting of control response efficacy', paper presented at the 31st Meeting of the Psychonomic Society, New Orleans (1990).

Harvey, N., Koehler, D. J. and Ayton, P., 'Actor–observer differences in judgmental probability forecasting of control response efficacy', manuscript in preparation, University College London and University of Waterloo (1996).

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

Hutchinson Pocket Factfinder (2nd edition), Oxford: Helicon, 1994.

Kelley, C. M. and Lindsay, D. S., 'Remembering mistaken as knowing: ease of retrieval as a basis for confidence in answers to general knowledge questions', *Journal of Memory and Language*, **32** (1993), 1–24.

Koehler, D. J., 'Hypothesis generation and confidence in judgment', *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **20** (1994), 461–9.

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

Lichtenstein, S., Fischhoff, B. and Phillips, L. D., 'Calibration of probabilities: The state of the art to 1980', in Kahneman, D., Slovic, P. and Tversky, A. (eds), *Judgment under Uncertainty: Heuristics and biases*, (pp. 306–34), Cambridge: Cambridge University Press, 1982.

Lord, C. G., Lepper, M. R. and Preston, E., 'Considering the opposite: a corrective strategy for social judgment', *Journal of Personality and Social Psychology*, **47** (1984), 1231–43.

Mehle, T., Gettys, C. F., Manning, C., Baca, S. and Fisher, S., 'The availability explanation of excessive plausibility assessments', *Acta Psychologica*, **49** (1981), 127–40.

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

Slovic, P. and Fischhoff, B., 'On the psychology of experimental surprises', *Journal of Experimental Psychology: Human Perception and Performance*, **3** (1977), 544–551.

Tetlock, P. E. and Boettger, R., 'Accountability amplifies the status quo effect when change creates victims', *Journal of Behavioral Decision Making*, **7** (1994), 1–23.

Tetlock, P. E., Skitka, L. and Boettger, R., 'Social and cognitive strategies for coping with accountability: Conformity, complexity, and bolstering', *Journal of Personality and Social Psychology*, **57** (1989). 632–40.

Tversky, A. and Koehler, D. J., 'Support theory: a nonextensional representation of subjective probability', *Psychological Review*, **101** (1994), 547–67.

Yates; J. F., *Judgment and Decision Making*, Englewood Cliffs, NJ: Prentice Hall, 1990.

Authors' biographies:
Derek Koehler is an assistant professor in the Department of Psychology at the University of Waterloo. His research interests include probability judgment and social cognition.

Nigel Harvey is a Reader in Experimental Psychology at University College London. His research interests include judgmental forecasting and control of dynamical system behavior.

Authors' addresses:
Derek Koehler, Department of Psychology, University of Waterloo, Waterloo, Ontario, N2L 3G1, Canada.

Nigel Harvey, Department of Psychology, University College London, Gower Street, London WC1E 6BT, UK.

