Intuitive Evaluation of Likelihood Judgment Producers: Evidence for a Confidence Heuristic

PAUL C. PRICE, California State University, Fresno, USA

ERIC R. STONE, Wake Forest University, USA

ABSTRACT

This research tests the hypothesis of Yates et al. (1996) that people prefer judgment producers who make extreme confidence judgments. In each of three experiments, col-lege students evaluated two fictional financial advisors who judged the likelihood that each of several stocks would increase in value. One of the advisors (the moderate advi-sor) was reasonably well calibrated and the other (the extreme advisor) was overconfi-dent. In all three experiments, participants tended to prefer the extreme advisor. Experiments 2 and 3 showed that the advisors' confidence influenced participants' per-ception of their knowledge, and Experiment 3 showed that it influenced their perception of the number of categorically correct judgments they made. Both of these variables were, in turn, related to participants' preferences. Experiment 3 also suggested that need for cognition and right-wing authoritarianism are positively related to preference for the extreme advisor. A quantitative model is presented, which captures the basic pattern of results. This model includes the assumption that people use a confidence heuristic; they assume that a more confident advisor makes more categorically correct judgments and is more knowledgeable.

KEY WORDS

confidence judgments; calibration; advisors; individual differences

Likelihood judgment has been a focal topic of behavioral decision research since the inception of the field. Most of this research has focused on either the objective accuracy of people's likelihood judgments or the cognitive processes underlying them (see, e.g., Wright & Ayton, 1994). Recently, however, there has been an increasing appreciation of the fact that likelihood judgments are often made by one person to be used by another. This happens, for example, whenever a physician, attorney, or other expert makes a likelihood judgment that enters into the decision process of a client. Referring to the former as *judgment producers* and the latter as *judgment consumers*, Yates et al. (1996) have argued that it is as important to understand how judgment consumers perceive, evaluate, and use likelihood judgments as it is to understand how judgment producers make them (see also Harvey, Harries, & Fischer, 2000; Keren, 1997; Keren &

Teigen, 2001; Sniezek & Buckley, 1995; Sniezek & Von Swol, 2001; Yaniv & Foster, 1995; Zarnoth & Sniezek, 1996).

The present research concerns a particular likelihood judgment phenomenon, overconfidence, from this perspective. It is now well established that under many conditions people's explicit judgments of the likelihood that some target event will occur are more extreme than is warranted by the relative frequency with which the target event actually occurs (e.g., Lichtenstein, Fischhoff, & Phillips, 1982; McClelland & Bolger, 1994; Yates, 1990). Previous research has focused extensively on the conditions under which overconfidence occurs, the psychological processes that produce it, and methods for reducing it. Very little research, however, has concerned overconfidence from the judgment consumer's perspective (Keren, 1997). Furthermore, the research that has been published on this topic suggests that overconfidence is not nearly as important to judgment consumers as it is to judgment researchers. In fact, there is reason to believe that people are more sensitive to the absolute level of confidence exhibited by a judgment producer than to the match between his or her confidence and the relative frequency of the target event (Yates et al., 1996). For this reason, we propose here that people use a *confidence heuristic*, according to which they use a judgment producer's confidence as a cue to his or her knowledge, competence, or correctness.[1] In the remainder of this article, we first review research that suggests this hypothesis and then report three experiments that support it. Finally, we present a simple quantitative model that incorporates this assumption and accounts for our results.

PREVIOUS RESEARCH

Consider first the research of Sniezek and colleagues. Zarnoth and Sniezek (1996) asked participants to respond to several different kinds of items (e.g., math problems, analogy problems, forecasting problems) and to express their degree of confidence that they had responded correctly to each one. Participants did this first as individuals and again in small groups. In neither case did they receive feedback about the correct responses. Zarnoth and Sniezek found that the group responses tended to match the individual responses of the most confident group members, even when the individual responses of the most confident group members were incorrect. More recently, Sniezek and Von Swol (2001) found that judges were more apt to take advice that was expressed with high, rather than low, confidence. An interesting parallel to these results is found in research on eyewitness testimony. This research has shown that the perceived confidence of an eyewitness is the single best predictor of his or her perceived credibility (Whitley & Greenberg, 1986), even though eyewitness confidence is only weakly related to eyewitness accuracy (Wells & Murray, 1984). Both sets of results suggest the use of a confidence heuristic, according to which people assume that the most confident individuals are the most likely to be correct.

One feature of the work by Sniezek and colleagues (Sniezek & Von Swol, 2001; Zarnoth & Sniezek, 1996)—which is also generally true in eyewitness testimony situations—is that participants evaluated confidence judgments with little or no information about the match between these judgments and the true state of the world. Note that in such situations, using a confidence heuristic is not unreasonable. Consider two judgment producers, A and B. Judgment Producer A is very confident that proposition P is true and Judgment Producer B is only slightly confident that the converse of P is true. To the extent that there is a positive relationship between confidence and accuracy across judgment producers—no matter how weak—Judgment Producer A is more likely to be correct. In this type of situation, then, it may reasonable to give more credence to a confident judgment producer than to a less confident one. Indeed, Sniezek and Von Swol found that when judgment consumers relied on advice expressed with greater confidence, their own judgmental accuracy was better.

---

[1]By 'correctness,' we mean categorical correctness. That is, if the advisor were to make categorical judgments about the occurrence or non-occurrence of the target event, how many would be correct?

In many other situations, however, the judgment consumer may have information about the judgment producer's past performance. In particular, the judgment consumer may be able to evaluate each of a series of judgments made by the same judgment producer, each time receiving feedback about the true state of the world. In this kind of situation, whether or not a judgment producer is overconfident can, in principle, be detected. Imagine, for example, that one's personal physician or financial advisor repeatedly and confidently makes diagnoses or stock price forecasts that turn out to be incorrect. What role does confidence, and the match between confidence and the true state of the world, play in the presence of such outcome feedback? Do people still use a confidence heuristic?

The work of Yates et al. (1996) suggests an answer to this question. They offered people a choice between two judgment producers as potential advisors. One of the judgment producers was well calibrated. That is, this judgment producer's likelihood judgments matched the relative frequency of occurrence of the target event. The second judgment producer was not well calibrated but was better able to distinguish occasions on which the target event would occur from occasions on which it would not. That is, the second judgment producer had better *discrimination* (also known as resolution). The second judgment producer was also more likely to make judgments greater than 50% when the target event occurred and less than 50% when it did not. Each probability judgment, and the associated outcome (i.e., whether or not the target event actually occurred), was printed on a single index card. Thus, the probability judgments of the two advisors were represented as two decks of 48 cards each. Participants studied both decks of cards, sorting or arranging them in any way that was helpful to them, and then chose the advisor that they would prefer to hire. In one condition, the advisors were described as financial advisors who judged the probability that each of 48 stocks would increase in value, and in another condition, they were described as meteorologists who judged the probability that it would rain on each of 48 consecutive days.

The primary result was that 28 out of 36 participants (78%) preferred the judgment producer with better discrimination to the one with better calibration. The most straightforward interpretation of this result is that judgment consumers are particularly concerned with either the discrimination ability of judgment producers or with their correctness. It is important to note, however, that the judgment producer with better discrimination was also overconfident. Furthermore, in analyzing the reasons participants gave for their choices, Yates et al. (1996) found that many of them reported that they preferred the judgment producer with better discrimination because he was extremely confident in his judgments. Many participants also referred to the correctness of the judgment producers, but not a single one mentioned anything that could be construed as calibration or discrimination. This suggests that judgment consumers may be particularly concerned with the absolute level of confidence of judgment producers, perhaps in addition to their correctness. It also suggests that they are not particularly concerned with the correspondence between judgment producers' likelihood judgments and the true state of the world.

The following three experiments were designed to explore this possibility. The first experiment demonstrates a general tendency to prefer an extreme, overconfident judgment producer to a more moderate one, holding constant both the discrimination ability and correctness of the two judgment producers. Experiments 2 and 3 provide evidence that this effect results from use of a confidence heuristic. Again, in the general discussion, we present a quantitative model that captures the basic pattern of results and suggests several interesting hypotheses for further study.

EXPERIMENT 1

The purpose of Experiment 1 was to test the hypothesis that people prefer an overconfident advisor to a well-calibrated one, even when discrimination and correctness are controlled. To do so, we created an experimental paradigm similar to that of Yates et al. (1996), in which participants studied the likelihood judgments of two potential financial advisors (along with information about the occurrence or non-occurrence of the target

event) and then stated a preference for one of the two advisors. The present experiments differed from those of Yates et al., however, in that participants were presented with the judgment–outcome pairs sequentially on a video display. This procedure was meant to simulate situations in which likelihood judgments and outcomes are experienced sequentially, and have not been carefully recorded and organized for the judgment consumer.

Method

Participants

The participants were 35 undergraduate students at the University of Michigan. They participated in return for partial credit in an introductory psychology course.

Stimulus data

We created two sets of likelihood judgments and associated outcomes, which were similar in their quantitative characteristics to those used by Yates et al. (1996).[2] We began by constructing a single stimulus data set of 24 cases, where each case consisted of a judgment of the probability that a target event would occur (expressed as a percentage from 0% to 100%), and an associated outcome (whether or not the target event occurred; see Appendix A.) The judgments, which ranged from 16 to 84%, had a mean of 50% and a standard deviation of 24%. The target event occurred for half the cases and failed to occur for half the cases. The judgments and outcomes in this *moderate data set* were paired such that the data set exhibited reasonably good calibration. We then constructed a second data set by adding 15% to each judgment above 50% and subtracting 15% from each judgment below 50% in the first data set. This *extreme data set*, therefore, exhibited considerable overconfidence but the same level of discrimination and number of categorically correct predictions as the moderate data set.

Before proceeding, it is worth elaborating on these last few points. First, with an externally defined target event (e.g., 'This stock will increase in value.') and a full-range (0 to 100%) response format, overconfidence refers to judgments above 50% that are too high compared to the relative frequency of the target event and judgments below 50% that are too low (Yates, 1990).[3] Given this conceptual definition, therefore, the extreme data set exhibited greater overconfidence than the moderate data set. One way to see this difference quantitatively is to compare the calibration indexes for the two data sets. The calibration index is a measure of the deviation of the judgments from perfect calibration (see Yates, 1994, for details).[4] The calibration index for the moderate data set (0.0347) was slightly lower (i.e., better) than the calibration index for the extreme data set (0.0472). Note, however, that the calibration index is not a pure measure of overconfidence because it reflects other forms of miscalibration as well (even underconfidence). For this reason, it is also useful to convert the likelihood judgments so that they concern an internally defined target event (e.g., 'I am correct') using a half-range (50 to 100%) response format (e.g., Ronis & Yates, 1987). To do so, one simply interprets judgments greater than 50% as categorical judgments that the target event will occur, with a judged

---

[2]Yates et al. (1996) used 48 judgment–outcome pairs. The mean likelihood judgment across the two data sets was approximately 45% and the target event occurred 42% of the time. The mean probability score exhibited by both data sets was approximately 0.1900.

[3]Note that the term 'overconfidence' has been used in two different ways, referring both to overestimating the relative frequency of the target event, and to providing probability judgments that are more extreme than is warranted by the actual categorical correctness of the judgment producer (see Lichtenstein et al., 1982). As the focus of our work is on the confidence of the judge, we are using the latter definition for this work.

[4]The computation of calibration and discrimination indexes generally requires rounding the likelihood judgments into a limited set of categories. Often, the likelihood judgments are rounded to the nearest tenth (e.g., judgments of 72% or 73% are rounded to 0.70; Yates, 1994). Here, however, we have rounded to the nearest twentieth (e.g., a judgment of 72% was rounded to 0.70 and a judgment of 73% was rounded to 0.75). Therefore, when 15% is added to or subtracted from each likelihood judgment in the moderate data set, the resulting likelihood judgment in the extreme data set is always more extreme by precisely three judgment categories. This eliminates differences in the calibration and discrimination indexes between the two data sets that are due solely to rounding error.

likelihood of being correct equal to the original judgment. Likewise, one interprets judgments less than 50% as categorical judgments that the target event will not occur, with a judged likelihood of being correct equal to 100% minus the original judgment (see, e.g., Stone & Opel, 2000). Given this procedure, an appropriate index of overconfidence is now mean confidence minus percentage correct (see Brenner et al., 1996), sometimes referred to as *bias* (Yates, 1994). After converting the stimulus data in this way, the bias statistic for the moderate data set was −3.33%, indicating slight underconfidence. The bias statistic for the extreme data set, however, was 11.67%, indicating considerable overconfidence.

Similarly, the way in which the data sets were constructed guarantees that the level of discrimination exhibited by them is identical. Again, discrimination refers to a judgment producer's ability to discriminate occasions on which the target event will occur from occasions on which it will not. As discussed by Yates (1994), this ability is unaffected by the specific numerical labels (e.g., 60% vs. 75%) given to the different judgment categories used. For this reason, making each likelihood judgment in the moderate data set more extreme by 15% produces an extreme data set exhibiting an identical level of discrimination. To make a quantitative comparison between the data sets, however, we can compute their discrimination indexes and show that they are equal (0.0972). A final point of interest is that in terms of the most commonly used index of overall likelihood judgment accuracy, the mean probability score, the moderate data set exhibited slightly greater accuracy (0.1875) than the extreme data set (0.2000). This is a straightforward implication of the fact that the moderate data set exhibited better calibration than the extreme data set, and that the two data sets were equal in terms of both discrimination and the relative frequency with which the target event occurred (again, see Yates, 1994, for details).

Design and procedure

The stimuli were presented, and participants' responses collected using personal computers. Participants worked in small, non-interacting groups, with each individual at a separate station. They began by reading the instructions reproduced in Appendix B, which described their task as that of deciding which of two financial advisors, Green or Brown, they would prefer to hire. Then, on each trial, the following information was presented on the computer monitor. First, there was one of two line-drawing images of a man in a business suit with a briefcase, which represented either Advisor Green or Advisor Brown. Second, there was the name of a fictional stock about which this financial advisor was said to have made a judgment. Third, there was the financial advisor's judgment of the probability that the stock would have increased in price at the end of three months (expressed as a percentage from 0% to 100%). Finally, there was a statement of whether or not the stock actually increased or decreased in price at the end of three months. Figure 1 shows the complete stimulus display for a single trial.

For half the participants, Advisor Green was associated with the moderate data set and Advisor Brown was associated with the extreme data set. For the rest of the participants, this pairing was reversed. The judgment–outcome pairs for Advisor Green were randomly intermixed with those for Advisor Brown, and each participant saw the 48 judgment–outcome pairs (24 for each advisor) in a different random order. Presentation of the trials was self-paced. Participants could study each judgment–outcome pair as long as they wanted, but they were not allowed to take written notes. At the end of the data presentation, participants were asked which of the two financial advisors they would prefer to hire.

Results and discussion

Of primary interest is that 25 out of the 35 participants (71%) preferred the extreme advisor to the moderate one. This percentage is significantly greater than 50%, as indicated by a two-tailed binomial test, $p = 0.02$.[5]

[5]All *p*-values reported in this article are two-tailed values.

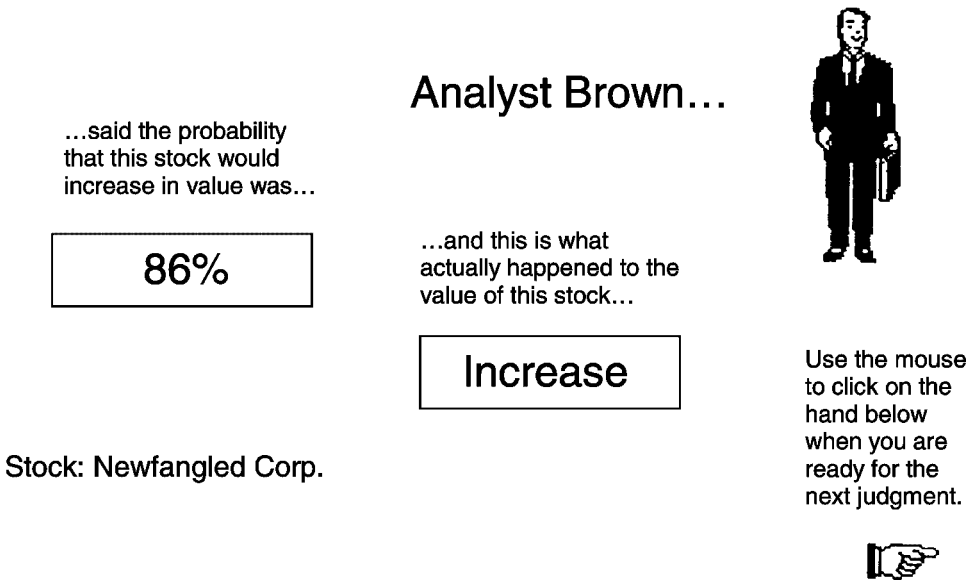

Figure 1. Sample stimulus display

In a sense, however, testing against a null hypothesis of equal preference understates the effect. According to traditional methods for assessing the accuracy of probability judgments (Yates, 1994), the moderate advisor was more accurate and might reasonably have been expected to be preferred by the majority of participants. The fact that our participants preferred an overconfident advisor to a better calibrated one, with other important aspects of accuracy held roughly constant, strongly suggests that many judgment consumers consider extreme confidence to indicate competence, knowledge, or correctness.

An alternative interpretation of the results of Experiment 1, however, is that participants preferred the extreme advisor because they found his judgments easier to discriminate. It may not have been his extremity per se, but rather the fact that participants were better able to perceive that his judgments differed from each other. For this reason, Experiments 2 and 3 included additional dependent variables to help determine whether participants do, in fact, consider the extreme advisor more knowledgeable than the moderate advisor, and to determine to what extent participants are able to process accurately information about the advisors' likelihood judgments.

EXPERIMENT 2

Experiment 2 was a replication of Experiment 1, but with some additional elements to provide deeper insight into the reasons for participants' preferences. Recall that the rationale for expecting the extreme advisor to be preferred to the moderate one was that people use a confidence heuristic. That is, people interpret greater confidence as indicating greater knowledge, competence, or correctness. If this is true, then we would expect participants who prefer the extreme advisor to consider him to be more knowledgeable. But what about the substantial minority of participants who prefer the moderate advisor? One possibility is that they interpret the extreme advisor's confidence as indicating that he is actually less knowledgeable. Another is that they also interpret the extreme advisor's confidence as indicating that he is more knowledgeable, but prefer the moderate advisor for some other reason, such as his seeming more honest in making his likelihood judgments.

To address these issues, we asked participants a number of additional questions at the completion of the experiment. Specifically, we asked which of the two financial advisors they thought was more knowledgeable. We also asked which of the two advisors they thought was more honest and more optimistic. Both of these characteristics were cited by the participants of Yates et al. (1996) as reasons for their preferences, and it seems plausible that they might determine participant preferences in the present experimental paradigm.

A final manipulation in Experiment 2 was for the financial advisors to be men for half the participants and women for the other half. This manipulation was included to examine whether or not extreme judgments would be interpreted as positively when made by women as by men. Demonstrating that the preference for the extreme advisor holds regardless of the gender of the advisors would be an important first step in generalizing this basic result across different kinds of advisors.

Method

Participants

The participants were 80 undergraduate students at Wake Forest University, who participated in partial fulfillment of an introductory psychology course requirement.

Design and procedure

This study was identical to Experiment 1, with the following exceptions. First, half the participants were presented with the images of male business executives used in Experiment 1, but the other half were presented with line-drawing images of female business executives. Second, after expressing their preference for one of the two financial advisors, participants were asked which one was more knowledgeable. They indicated their response by using the mouse to point and click on one of three icons labeled *Brown*, *Green*, and *Neither*. Then they were asked which financial analyst was more honest and which was more optimistic, responding to each question in the same way. The *Neither* response option was included for these questions because it seemed reasonable to think that some participants might not perceive a difference between the two advisors on these dimensions.

Results and discussion

Participants were equally likely to prefer the overconfident advisor regardless of whether the advisors were men or women. In the male-advisor condition, 25 out of 40 participants (63%) preferred to hire the extreme advisor, while in the female-advisor condition, 26 out of 40 participants (65%) preferred to hire the extreme advisor. Because there was essentially no difference between the male- and female-advisor conditions, the data were aggregated for all subsequent analyses. Overall, 51 out of 80 participants (64%) preferred to hire the overconfident advisor. This percentage was comparable to that from Experiment 1, and it was again significantly greater than 50%, as indicated by a two-tailed binomial test, $p = 0.02$.

Recall that in response to each of the three follow-up questions (e.g., regarding which advisor was more knowledgeable), participants chose either the extreme advisor, the moderate advisor, or neither. Table 1 presents the frequencies of each of these three responses to each of the three follow-up questions. Note that one of the participants did not respond to these questions, so that the total number of participants responding was 79. There was a strong association between which advisor participants preferred and which one they thought was more knowledgeable, $\chi^2(2) = 33.03$, $p < 0.001$. When participants had an opinion on which advisor was more knowledgeable, they tended to think it was their preferred advisor. There were no such associations, however, between which advisor participants preferred and which one they thought was more honest, $\chi^2(2) = 3.75$, $p = 0.15$, or which one they thought was more optimistic, $\chi^2(2) = 1.63$, $p = 0.44$.

To summarize, the majority of participants interpreted the extreme advisor's greater confidence as indicating greater knowledge and preferred that advisor. However, there was a sizable minority of participants

Table 1. Number of participants in Experiment 2 giving each possible response to questions about which advisor was more knowledgeable, honest, and optimistic ($N = 79$)

		Preferred advisor	
Question	Response	Extreme	Moderate
Which advisor was more knowledgeable?	Extreme	34	03
	Moderate	02	15
	Neither	14	11
Which advisor was more honest?	Extreme	13	03
	Moderate	10	10
	Neither	27	16
Which advisor was more optimistic?	Extreme	20	15
	Moderate	20	11
	Neither	10	03

who thought that the moderate advisor was more knowledgeable and tended to prefer that advisor. Although this result seems at first to be inconsistent with our hypothesis that people in general use a confidence heuristic, the quantitative model presented in the general discussion shows that it is not. It is also important to consider a possible artifactual explanation of these results, that participants first formed a preference for one advisor and then simply reported that their preferred advisor was more knowledgeable, perhaps as a justification for their preference or as the result of a halo effect (Cooper, 1981). If this were true, however, there should have been a parallel tendency for participants to report that their preferred advisor was also both more honest and more optimistic. The fact that there was no tendency for participants to report that their preferred advisor was either more honest or more optimistic suggests that they did not feel compelled to respond to the follow-up questions in whatever way justified their preference.

EXPERIMENT 3

We had two main goals in conducting Experiment 3. First, we wanted to test in a more direct way the hypothesis that participants use the advisors' relative confidence as a cue to their relative correctness. Specifically, we wanted to determine whether the advisors' confidence—and participants' perceptions of it—actually influenced participants' perceptions of the advisors' correctness and to see how these perceptions related to their preferences. Second, we wanted to explore the nature of the individual differences in advisor preference observed in Experiments 1 and 2. Specifically, we examined whether these differences could be explained in part by the personality measures of need for closure, need for cognition, and right-wing authoritarianism.

To help us achieve the first of these goals, we changed the format in which the advisors made their likelihood judgments. In Experiments 1 and 2, the advisors judged the likelihood of an externally defined target event using a full-range (0 to 100%) format. For Experiment 3, we modified this procedure so that the advisors judged the likelihood of an internally defined target event ('I am correct') using a half-range (50 to 100%) format. The primary rationale for this change is that it allowed us to ask participants to estimate the percentage of stocks for which each advisor was correct in his or her categorical judgment. Clearly, if participants use a confidence heuristic, they should estimate the extreme advisor's percentage correct to be greater than the moderate advisor's percentage correct.

A second advantage of switching to an internal target event and half-range format is that it allows us to eliminate one other potential explanation for the results of Experiments 1 and 2. Although calibration and discrimination are the two most frequently discussed aspects of likelihood judgment accuracy, an alternative perspective is provided by the covariance approach (see Yates, 1982; 1994). One aspect of accuracy within

this approach is *slope*, which is the difference in the mean probability judgment when the target event occurs and the mean probability judgment when the target event does not occur. In Experiments 1 and 2, the extreme advisor had a greater slope than the moderate advisor.[6] This is because, with an externally defined target event and a full-range response format, making uniformly more-extreme likelihood judgments must increase slope if the judgment producer has any positive level of discrimination. In other words, given the procedure used in the first two studies, judgment extremity and slope were *necessarily* confounded. Although we doubt that participants would spontaneously compute the measures necessary to calculate slope, it is theoretically possible that they were responding to a difference in this variable rather than to a difference in judgment extremity. Switching to an internal target event, however, eliminates this possibility because slope is identical for the two data sets (0.0222 or 2.22%).

The second goal of Experiment 3 was to examine the relationship between advisor preference and three well-known individual difference measures that seemed potentially relevant in the present context. The first is the need for closure (NFClo), which refers to a general tendency to prefer certain over uncertain knowledge (Kruglanski, 1989). According to Kruglanski, Webster, and Klem (1993), NFClo 'represents a desire for a clear-cut opinion on a judgmental topic' (p. 861). Therefore, it seems reasonable to think that people who are high in NFClo might be more likely to prefer the extreme advisor because he or she generally provides a more clear-cut opinion than does the moderate advisor. The second individual difference is need for cognition (NFCog), which refers to a general tendency to value and engage in effortful thought (Cacioppo et al., 1996). This variable seems relevant here in two ways. First, it might be that people who are low in NFCog are more likely to prefer the extreme advisor because his or her confidence judgments leave less room for doubt, and therefore less need for additional thinking, than do those of the moderate advisor. Second, the experimental task itself requires effortful thought, so NFCog might be related to the extent to which people process information about the advisors' judgments. Specifically, participants high in NFCog might be more likely to prefer the extreme advisor because they would be more apt to perceive the large difference in confidence between the two advisors.

Finally, we examined differences in right-wing authoritarianism (RWA), which refers to a cluster of social beliefs concerning deference to authority, authoritarian aggression, and conventionalism (Altmeier, 1981). It seems reasonable to believe that highly authoritarian people would be particularly inclined to favor extreme, highly confident judgments—at least in people in positions of authority. In support of this line of reasoning, Wright and Phillips (1976, as cited in Lichtenstein et al., 1982), in an investigation of the relationship between calibration and different personality measures, found significant (though modest) relationships only with authoritarianism. Participants high in RWA, then, should be particularly apt to prefer the extreme advisor due to the value placed on his high confidence.

Method

Participants

The participants were 82 undergraduate students (49 women and 33 men) at California State University, Fresno. They participated in partial fulfillment of an introductory psychology course.

Design and procedure

The design and procedure of Experiment 3 were very similar to those of Experiments 1 and 2. There were a number of minor differences in the way the stimuli were displayed, however, including the use of a different

[6]From the covariance perspective, the reason the extreme data set exhibited a greater (i.e., worse) mean probability score is that it exhibited greater scatter. After converting the judgments to refer to an internally defined target event using a half-range response format, both the slope and scatter of the two data sets were identical. Only the bias exhibited by the two data sets was different.

icon to represent each of the financial advisors (both male). In addition, the financial advisors were referred to as Advisors Blue and Green (rather than Brown and Green), and the icon representing each advisor was always displayed in the color corresponding to his name.

The primary difference between Experiment 3 and Experiments 1 and 2 was that instead of judging the probability that the stocks would increase in value, the financial advisors stated categorically whether they believed the stocks would increase or decrease in value and also judged the probability that they were correct in their categorical judgments. The stimulus data from Experiments 1 and 2 had to be altered, therefore, in the following way. For cases in which the original probability judgment was greater than 50%, the financial advisors stated that they believed the stock would increase in value and judged the probability that they were correct to be equal to the original probability judgment. For cases in which the original probability judgment was less than 50%, the financial advisors stated that they believed the stock would decrease in value with probablilty equal to 100% minus the original probability judgment. This procedure has been used previously to translate likelihood judgments pertaining to an external target event to an internal target event (e.g., Ronis & Yates, 1987). As in Experiments 1 and 2, information about whether each stock increased or decreased in value was also presented on each trial.

Participants began by reading a new set of instructions, which was a revised version of that from Experiments 1 and 2 reflecting the changes in procedure. After the presentation of the stimulus data, participants again stated which of the two advisors they preferred to hire. They also estimated each advisor's percentage of correct categorical judgments and average confidence judgment (in each case by typing a number from 0 to 100%). The order in which these estimates were made was as follows: Advisor Blue's percentage correct, Advisor Green's percentage correct, Advisor Blue's average confidence judgment, Advisor Green's average confidence judgment. Since each advisor was associated with the extreme data set half the time, half of the participants began by estimating the percentage correct of the extreme advisor and the other half by estimating the percentage correct of the moderate advisor. Finally, participants indicated which financial advisor they believed was more knowledgeable about the stock market. In this experiment, however, they were forced to select one advisor or the other; *Neither* was not an option.

Results and discussion

For the remainder of this section, we use the following presentation plan. First, we present the results regarding our primary dependent variable: preference for the extreme versus moderate advisor. Second, we describe the results concerning the secondary dependent variables of perceived knowledge, perceived confidence, perceived percentage correct, and perceived bias or overconfidence. Third, we examine whether the pattern of results concerning the secondary dependent measures holds for all participants regardless of their preference. And finally, we describe the role played by the three individual difference measures included in this study.

Four participants gave either average-confidence or percentage-correct estimates that were far removed from the actual stimulus data. All four of them made one or more estimates that were less than 50% (three of them made estimates of 20% or less), which strongly suggests that they misunderstood the task. The following results, therefore, are based on data from the other 78 participants.

Advisor preference

As in Experiments 1 and 2, the majority of participants preferred the extreme advisor to the moderate advisor. Specifically, 49 of the 78 participants (63%) preferred the extreme advisor. An exact binomial test revealed that this result would be unlikely if the percentage of the population preferring each advisor were 50% ($p = 0.03$). A somewhat greater percentage of men (24 out of 33) than women (25 out of 45) preferred the extreme advisor, but this difference was not statistically significant, $\chi^2(1) = 2.40$, $p = 0.12$. Aggregating

Table 2. Mean average-confidence, percentage-correct, and implicit bias estimates (with actual values) for the two advisors in Experiment 3 across all participants ($N = 78$)

	Advisor			
	Moderate		Extreme	
Estimate	Actual value	Estimated value	Actual value	Estimated value
Average-confidence	71.67%	70.38 (8.04)	86.67%	83.42 (7.88)
Percentage-correct	75.00%	73.82 (8.06)	75.00%	76.56 (8.89)
Implicit bias	−3.33%	−3.44 (9.03)	11.67%	6.86 (10.03)

Implicit bias estimates are the difference between participants' average-confidence and percentage-correct estimates. Standard deviations are in parentheses.

over men and women, it is evident that the overall percentage of participants preferring the extreme advisor was very similar to the percentages from Experiments 1 and 2, despite the changes in design and procedure. That is, the percentage of participants preferring the extreme advisor was about the same even though the advisors' judgments were based on a two-stage procedure with an internal target event and half-range response format. Because all other measures of likelihood judgment accuracy were controlled, it is clear that the extremity of the advisors' judgments was the crucial independent variable.

Perceived knowledge, average confidence, and percentage correct

As in Experiment 2, the majority of participants thought that the extreme advisor was more knowledgeable than the moderate advisor. Specifically, 47 of the 78 participants (60%) thought the extreme advisor was more knowledgeable, although an exact binomial test provided only marginal evidence that this percentage was significantly different from 50%, $p = 0.09$.

Recall that the extreme advisor's average confidence was 86.67%, the moderate advisor's average confidence was 71.67%, and both advisors were correct 75% of the time. As shown in Table 2, participants correctly perceived that the extreme advisor was more confident than the moderate advisor. This difference was statistically significant, $t(77) = 9.60$, $p < 0.001$. In support of the idea that participants used the advisors' confidence as a cue to their correctness, participants tended to overestimate the percentage correct of the extreme advisor and underestimate it for the moderate advisor. This difference was also statistically significant, $t(77) = 2.31$, $p = 0.02$. Finally, we computed a pair of implicit-bias estimates for each participant by subtracting his or her percentage-correct estimate from his or her average-confidence estimate for each advisor. This analysis showed that, in one sense, participants did correctly perceived the extreme advisor to be more overconfident than the moderate advisor, $t(77) = 7.13$, $p < 0.001$.

The data regarding perceived average confidence and perceived percentage correct also help clarify what is driving participants' perceptions of knowledge. Specifically, we were interested in determining whether knowledge was equated more with high confidence, a high percentage correct, or a lack of bias.[7] To explore this issue, we computed a new set of dependent variables, each of which was the difference between participants' estimates for the extreme and moderate advisor. Specifically, we subtracted each participant's average-confidence estimate for the moderate advisor from his or her average-confidence estimate for the extreme advisor, such that a difference of 15 would accurately reflect the stimulus data. We then followed an analogous procedure for the percentage-correct and bias estimates, thus creating variables reflecting perceived differences in confidence, percentage correct, and bias between each of the advisors.

[7]We thank an anonymous reviewer for suggesting this possibility.

Table 3. Differences in average-confidence, percentage-correct, and implicit bias estimates between the extreme and moderate advisors, as a function of the advisor perceived as more knowledgeable and the preferred advisor

	Advisor judged more knowledgeable		Preferred advisor	
Estimate	Extreme ($n = 47$)	Moderate ($n = 31$)	Extreme ($n = 49$)	Moderate ($n = 29$)
Average-confidence	17.19 (9.69)	6.74 (12.54)	16.98 (10.73)	6.38 (11.19)
Percentage-correct	6.68 (9.23)	−3.22 (9.49)	7.59 (8.44)	−5.45 (8.32)
Implicit bias	10.51 (12.88)	9.97 (12.77)	9.39 (13.93)	11.83 (10.52)

Implicit bias estimates are the difference between participants average-confidence and percentage-correct estimates.
Standard deviations in parentheses.

As shown in Table 3, participants who said the extreme advisor was more knowledgeable perceived a greater difference in confidence ($M = 17.19$) than did participants who thought the moderate advisor was more knowledgeable ($M = 6.74$), $t(76) = 4.14$, $p < 0.001$. Also, participants who saw the extreme advisor as more knowledgeable perceived a greater correctness advantage ($M = 6.68$) for the extreme advisor than did participants who said the moderate advisor was more knowledgeable ($M = -3.22$), $t(76) = 4.59$, $p < 0.001$. However, there were no differences between the two sets of participants in terms of their implicit bias judgments, $t(76) = 0.18$, $p = 0.86$. An alternative approach to this issue is to compute point-biserial correlations between the knowledge judgment and each of the perceived difference measures. Doing so shows that which advisor participants thought was more knowledgeable was correlated 0.43 with perceived confidence, 0.47 with perceived correctness, but only 0.02 with perceived bias. Thus, participants' choice of which advisor was more knowledgeable appears to have been driven primarily by their perception of the advisors' confidence and percentage correct. There is no evidence that other factors, such as the extent of their overconfidence, had any role in the knowledge judgment.

Participants preferring the extreme advisor vs. participants preferring the moderate advisor

As in Experiment 2, the general tendency to see the extreme advisor as being more knowledgeable than the moderate advisor did not hold for those participants who preferred the moderate advisor. Specifically, 39 out of the 49 participants who preferred the extreme advisor (80%) thought that he was more knowledgeable, but 21 of the 29 participants who preferred the moderate advisor (72%) thought that he was more knowledgeable, $\chi^2\ (1) = 20.6$, $p < 0.001$.

Participants perceived correctly that the extreme advisor was more confident than the moderate advisor regardless of which advisor they preferred (see Table 3). However, the perceived difference in confidence between the advisors was greater for those who preferred the extreme advisor ($M = 16.98$) than for those who preferred the moderate advisor ($M = 6.38$), $t(76) = 4.15$, $p < 0.001$. Similar to the results regarding knowledge, participants who preferred the extreme advisor ($M = 7.59$) perceived a greater correctness advantage for the extreme advisor than did those who preferred the moderate advisor ($M = -5.45$), $t(76) = 6.63$, $p < 0.001$. There was, however, no difference in perceived bias depending on advisor preference, $t(76) = 0.81$, $p = 0.42$.

Finally, we examined whether the knowledge judgments might depend on different factors according to which advisor was preferred. For example, perhaps those who preferred the extreme advisor thought he was more knowledgeable because he was highly confident, while those who preferred the moderate advisor did so because he was less biased. To examine this question, we tested for interactions between preference and knowledge on each of the difference measures (perceived confidence, percentage correct, and overconfidence). In none of these cases did the interaction even approach significance (all Fs <1). Thus, it appears that all participants—even those who preferred the moderate advisor—associate knowledge with high confidence and a high percentage correct, rather than with other factors.

Individual difference measures

The previous analyses showed that preference and perceived knowledge appear to be a function of both perceived average confidence and perceived percentage correct (resulting in part from high perceived confidence). Specifically, it appears that the majority of the participants perceived a large difference in confidence between the two advisors, translated that difference into a perceived difference in percentage correct, and thus preferred the more extreme advisor. Nonetheless, there was a sizeable minority who preferred the moderate advisor. If the previous account is correct, then, it seems likely that individual differences in preference should arise from: (1) individual differences in an ability to correctly perceive the difference in confidence; and (2) individual differences in a tendency to translate that perceived difference in confidence into a perceived difference in percentage correct. The following analyses assume this framework.

Need for closure. There were no significant relationships between need for closure and participants' advisor preference, average-confidence estimates, or percentage-correct estimates (all ps >0.50).

Need for cognition. Participants who preferred the extreme advisor were higher in NFCog ($M = 3.49$, $SD = 0.41$) than were participants who preferred the moderate advisor ($M = 3.34$, $SD = 0.37$), although this difference was not statistically significant, $t(76) = 1.62$, $p = 0.11$. Recall that we hypothesized that NFCog might be positively related to preference for the extreme judge, under the assumption that only those participants high in NFCog would correctly perceive the difference in confidence between the two advisors. An examination of the relationship between NFCog and the difference variables described above supports this possibility. Specifically, there was a moderate relationship between NFCog and the perceived difference in confidence, $r = 0.21$, which was marginally statistically significant, $p = 0.06$. Dividing participants into three equal-sized groups (low, medium, and high in NFCog) helps to describe the effect more precisely. Specifically, participants low in NFCog perceived a much smaller confidence difference ($M = 9.65$) than did participants either medium ($M = 14.15$) or high ($M = 15.31$) in NFCog. It is worth emphasizing that participants who were medium and high in NFCog did a remarkably good job of estimating the confidence difference of 15% between the two advisors; it was only the group low in NFCog that seriously underestimated the difference. Note, however, that the relationship between NFCog and the perceived difference in percentage correct was much weaker, $r = 0.09$, $p = 0.44$. Thus, it appears that NFCog may be related to the tendency to perceive (accurately) the large difference in confidence between the two advisors, but not to the tendency to translate this perceived difference in confidence into a perceived difference in correctness.

Right-wing authoritarianism. Participants who preferred the extreme advisor had higher RWA scores ($M = 3.97$, $SD = 0.60$) than did those who preferred the moderate advisor ($M = 3.76$, $SD = 0.44$), although again this difference was not quite statistically significant, $t\,(76) = 1.61$, $p = 0.11$. In this case, however, any potential RWA-preference relationship appears not to be mediated by perception of confidence, but instead by perception of correctness. Specifically, participants higher in RWA perceived a greater difference in correctness than did those lower in RWA, $r = 0.25$, $p = 0.03$. This effect appears quite striking when the participants are divided according to whether they were low, medium, or high in RWA. Participants low in RWA ($M = -0.50$) and those medium in RWA ($M = 0.93$) perceived the two advisors as being correct equally often. Those high in RWA, however, perceived a very large difference in percentage correct ($M = 8.37$). There was, however, only a small, non-significant relationship between RWA and the perceived difference in confidence between the two advisors, $r = 0.12$, $p = 0.28$. Thus, these results suggest that being high in RWA is not related to the ability to perceive the difference in confidence between the two advisors, but that it may be related to the tendency to translate this perceived difference in confidence into a perceived difference in correctness.

Summary. These ideas suggest that the tendency to prefer the extreme advisor should be particularly strong among participants who are both high in NFCog (and therefore perceive the large difference in confidence)

and high in RWA (and therefore translate this perceived difference in confidence into a perceived difference in accuracy). Just such a pattern of results emerged when we performed median splits on both individual difference measures and categorized participants as low–low, low–high, high–low, and high–high (in NFCog and RWA, respectively). The percentage of participants who preferred the extreme advisor was 53.6 across the first three categories and 86.4 in the fourth. This difference was statistically significant, $\chi^2(1, N = 78) = 7.27$, $p = 0.007$.

GENERAL DISCUSSION

In each of the three experiments, there was a tendency for participants to prefer an extreme, overconfident advisor to a better-calibrated one. These results clearly support the idea that judgment consumers do not necessarily evaluate judgment producers in ways that are consistent with formal analyses of likelihood judgment accuracy (Yates et al., 1996). We have hypothesized that this is because people use a confidence heuristic; they rely on a judgment producer's confidence as a cue to his or her knowledge, competence, and correctness. This idea was supported, in particular, by the fact that in Experiment 3 there was a strong tendency for participants who perceived a greater difference in confidence between the two advisors to perceive the extreme one as more knowledgeable and correct more often, and to prefer him.

At first glance, however, this does not seem to explain why a substantial minority of participants perceived the extreme advisor to be somewhat more confident than the moderate advisor yet perceived him to be correct less often (and did not prefer him). One possibility is that only some participants used a confidence heuristic. Those who did tended to prefer the extreme advisor and those who did not tended to prefer the moderate advisor. However, it is important to realize that the pattern of results we found is also consistent with the possibility that all participants used a confidence heuristic. Specifically, it can be explained solely by positing a certain amount of random error in the judgment process. As several researchers have shown recently, random error can have substantial and often non-intuitive effects on judgment (Dougherty, Gettys, & Ogden, 1999; Erev, Wallsten, & Budescu, 1994; Juslin, Olson, & Björkman, 1997).

Consider a simple model in which participants prefer the advisor they perceive to have made the greater percentage of correct categorical judgments (and to consider that advisor the more knowledgeable of the two). Let us assume further that participants correctly perceive the difference in percentage correct on average, but with a certain amount of random error. The following equation captures these assumptions.

$$\Delta P_{\text{corr}} = (d_1 - d_2) + e_{\text{corr}} \tag{1}$$

In Equation 1, ΔP_{corr} is the perceived difference in percentage correct between the two advisors, d_1 is the percentage correct of Advisor 1, and d_2 is the percentage correct of Advisor 2. The error term, e_{corr}, is assumed to be normally distributed with a mean of 0 and a standard deviation of s_{corr}. It is assumed to reflect random error in the perception of both advisors' percentages correct and the computation of the difference between them. Finally, we assume that Advisor 1 is preferred to Advisor 2 when ΔP_{corr} is positive, but that Advisor 2 is preferred to Advisor 1 when ΔP_{corr} is negative.

It is clear, however, that Equation 1 is not sufficient to explain the entire pattern of results found in our studies. According to Equation 1, the fact that both advisors made correct categorical judgments for 75% of the stocks implies that ΔP_{corr} should have had a mean of 0. As a result, about 50% of the participants should have preferred each advisor. To account for this discrepancy, let us assume that people pay attention not only to the advisors' correctness but also to their confidence. Let us define the perceived difference in average confidence between the two advisors as follows:

$$\Delta P_{\text{conf}} = (f_1 - f_2) + e_{\text{conf}} \tag{2}$$

In Equation 2, ΔP_{conf} is the perceived difference in average confidence, f_1 is the average confidence of Advisor 1, f_2 is the average confidence of Advisor 2, and e_{conf} is again a normally distributed error term with a mean of 0 and a standard deviation of s_{conf}. Now, let us add to Equation 1 a correctness advantage, conferred on the more confident advisor, that is proportional to the perceived difference in average confidence. Combining the two error terms from Equations 1 and 2 for simplicity produces the following equation:

$$\Delta P_{corr} = (d_1 - d_2) + w\Delta P_{conf} + e \quad (3)$$

In Equation 3, w is a weighting parameter ($w > 0$), such that greater values of w lead to a greater correctness advantage for the more confident advisor, and e refers to the combined error term.

Equation 3 provides an excellent description of our results. Recall that the empirical mean of ΔP_{corr} across all participants (the difference between participants' percentage-correct estimates for the two advisors) was 2.74. Because the actual difference in the stimulus data ($d_1 - d_2$) was zero, 2.74 must represent the correctness advantage conferred on the extreme advisor: w (ΔP_{conf}). Furthermore, because the empirical mean of ΔP_{conf} across all participants was 13.03, w can be estimated to be 0.21. This means that for every increase of one percentage point in the perceived difference in confidence between the two advisors, the extreme advisor gained a correctness advantage of 0.21 percentage points. Of particular interest is that dividing the empirical mean of ΔP_{corr} (2.74) by its empirical standard deviation (10.48) results in a z-score of 0.26, which implies that ΔP_{corr} was positive for 60% of the participants. This is quite close to the percentage that preferred the extreme advisor in Experiment 3. Again, it is important to emphasize that this result is implied by a model that assumes that all participants use the confidence heuristic in the same way. The fact that a substantial minority of participants preferred the less confident advisor is a result of random error in the judgment process.

In addition, this model has the potential to integrate the suggestive results regarding NFCog and RWA. First, recall that participants who were higher in NFCog tended to perceive a greater difference in confidence between the two advisors. According to the model presented here, this implies that these participants would also tend to confer a greater correctness advantage on the extreme advisor, which would explain their greater tendency to prefer him. Second, recall that participants who were higher in RWA tended to perceive a greater difference in correctness between the two advisors. This could be because the parameter w is an increasing function of RWA, such that participants higher in RWA also tend to confer a greater correctness advantage to the extreme advisor, which would explain their greater tendency to prefer him. Specifically, we estimated the values of w from the data in the way described above, and found that these values were –0.04, 0.07, and 0.64 for participants low, medium, and high in RWA, respectively. Taken at face value, these results suggest that perhaps only participants high in RWA use the confidence heuristic to any significant extent, but more research is needed before we can conclude that with any definitiveness.

More generally, future research on the evaluation of likelihood judgment producers should test this basic model and its implications more thoroughly. One priority is to replicate the basic result across a range of judgment contexts and across judgment producers varying in their levels of confidence and correctness. One can then ask how the correctness advantage—and therefore advisor preference—varies as a function of these task characteristics, in addition to how it varies as a function of participant characteristics.

It is important to note here that much recent research on confidence and overconfidence has taken an ecological perspective. This perspective assumes that, with experience, people develop accurate representations of event frequencies and relative frequencies (e.g., Gigerenzer, Hoffrage, & Kleinbölting, 1991). An experienced financial advisor, for example, would have an accurate representation of the relative frequency with which stocks of various types increase and decrease in value over a given period. From this perspective, experienced judgment producers are not truly overconfident, in the sense that they believe they know more than they actually do. Instead, their apparent overconfidence is the result of the non-representative sampling of cases (e.g., a group of unusual stocks; Juslin, 1994) or of random error in the judgment process (Juslin,

Olsson, & Björkman, 1997; Soll, 1996). Although the present theoretical analysis is neutral with respect to the causes of overconfidence, a consideration of the ecological perspective raises at least two interesting issues.

First, the ecological perspective in general suggests that, with extensive experience, judgment consumers might come to learn whether a particular advisor's likelihood judgments are well calibrated. One might even conceptualize our participant's task as a single-cue learning task (see, e.g., Brehmer, 1988; Klayman, 1988; Slovic & Lichtenstein, 1971) in which they must learn the relationship between the advisors' confidence and correctness. From this perspective it seems quite plausible that a total of 48 trials was not enough for significant learning to occur. Perhaps if participants had seen 240 or 2400 they would have perceived that the extreme advisor was overconfident and tended to prefer the moderate advisor. Of course, there is no way to know this without actually conducting a much longer experiment. It is also worth pointing out that people rarely get that many opportunities to evaluate individual likelihood judgment producers in significant contexts. Consultations with financial advisors, lawyers, and physicians are not everyday occurrences. Moreover, given research on the phenomenon of illusory correlation (e.g., Chapman & Chapman, 1969), it is quite possible that even in the face of considerable evidence to the contrary, participants will maintain their belief that greater confidence implies greater correctness.

Second, the ecological perspective suggests that the confidence heuristic itself might be based on a roughly accurate generalization across different judgment producers in many contexts. In fact, we suspect that this is the case. People rightly believe that more confident judgment producers tend to be correct more often. What really is at issue is the extent to which people will rely on the cue of confidence to evaluate judgment producers in the presence (and absence) of other relevant information. We suggest here that people may rely on it quite heavily. It may even dominate more traditional measures of likelihood judgment accuracy like calibration, bias, and so forth.

A final point is that if judgment consumers do use a confidence heuristic in evaluating judgment producers, then this might help to explain the prevalence of overconfidence. It seems reasonably clear that many factors contribute to overconfidence, including the ecological structure of the task, random error in the judgment process, the lack of prompt and unambiguous feedback, lack of repeated trials under similar conditions, and lack of experience translating feelings of uncertainty into probability judgments (Arkes et al., 1987; Ayton, 1992; Einhorn, 1982; Fischhoff, 1982; McClelland & Bolger, 1994; Murphy & Brown, 1984). Another factor may simply be that making extreme likelihood judgments is impressive to others. It can even influence others' perception of how often one is correct and, therefore, how knowledgeable and competent one is. This may be one reason that expert judgment producers typically express extreme confidence in their judgment (Katz, 1984; Shanteau, 1988). It is becoming increasingly clear that a complete understanding of confidence judgments must include this social dimension.

APPENDIX A: STIMULUS DATA SETS

	Probability judgment	
Target event occurred?	Data set 1: Moderate	Data set 2: Extreme
Yes	84%	99%
Yes	82%	97%
Yes	78%	93%
Yes	76%	91%
Yes	73%	88%
Yes	71%	86%
Yes	67%	82%

Continues

APPENDIX A. CONTINUED

Yes	62%	77%
Yes	58%	73%
Yes	40%	25%
Yes	29%	14%
Yes	20%	05%
No	80%	95%
No	69%	84%
No	60%	75%
No	42%	27%
No	38%	23%
No	33%	18%
No	31%	16%
No	27%	12%
No	24%	9%
No	22%	7%
No	18%	3%
No	16%	1%

APPENDIX B: INSTRUCTIONS FOR EXPERIMENTS 1 AND 2

Imagine that you have inherited some money, and that some friends have recommended you hire a financial advisor to help you invest that money. Some of your friends, however, have recommended Advisor Green, whereas others have recommended Advisor Brown.

To decide which financial advisor to hire, you hold a competition in which each advisor judges the likelihood that several stocks will increase in value. Specifically, each advisor will judge the probability that each of 24 stocks will have increased in value at the end of three months. You will then check these judgments against the actual performance of the stocks and hire the advisor whose judgments you prefer.

The rest of this experiment, then, will consist of a number of 'trials.' On each trial, you will be presented with a judgment made by one analyst (Green or Brown) concerning one stock, and you will be presented with information about the actual performance of that stock. For example, on one trial you might be informed that Advisor Green (or Brown) said there was a 90% chance that Stock X would increase in value, and that Stock X did increase in value. On another trial you might be told that Advisor Green (or Brown) said there was a 20% chance that Stock Y would increase in value, and that Stock Y did not increase in value.

On each trial you will be presented with information about a likelihood judgment made by one of the two advisors, but on any given trial it might be a judgment made by Advisor Green or by Advisor Brown. Pay close attention to the performance of *each* advisor. Also, you may examine the judgments at your own pace, but you may not go back to examine a judgment once you have moved on to the next one. So again, pay close attention.

Note that the advisors' judgments (i.e., the percentages) do not indicate anything about how *much* the stocks are likely to increase. They only indicate what the advisors believed to be the likelihood that the stocks would increase in value (as opposed to decrease).

Once you have been presented with all the judgments of both advisors (48 total), you will be asked to indicate which advisor you would prefer to hire.

REFERENCES

Altmeier, B. (1991). *Right-wing authoritarianism*. Winnipeg: University of Winnipeg Press.

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

Ayton, P. (1992). On the competence and incompetence of experts. In G. Wright, & F. Bolger (Eds.), *Expertise and decision support* (pp. 77–105). New York: Plenum Press.

Brehmer, B. (1988). The development of social judgment theory. In B. Brehmer, & C. R. B. Joyce (Eds.), *Human judgment: The SJT view* (pp. 13–40). Amsterdam: Elsevier.

Brenner, L. A., Koehler, D. J., Liberman, V., & Tversky, A. (1996). Overconfidence in probability and frequency judgments: a critical examination. *Organizational Behavior and Human Decision Processes, 65*, 212–219.

Cacioppo, J. T., Petty, R. E., Feinstein, J. A., Jarvis, W., & Blair, G. (1996). Dispositional differences in cognitive motivation: the life and times of individuals varying in need for cognition. *Psychological Bulletin, 119*, 197–253.

Chapman, L. J., & Chapman, J. P. (1969). Illusory correlation as an obstacle to the valid use of psychodiagnostic signs. *Journal of Abnormal Psychology, 74*, 271–280.

Cooper, W. H. (1981). Ubiquitous halo. *Psychological Bulletin, 90*, 218–244.

Dougherty, M. R. P., Gettys, C. F., & Ogden, E. E. (1999). MINERVA-DM: a memory processes model for judgments of likelihood. *Psychological Review, 106*, 180–209.

Einhorn, H. J. (1982). Learning from experience and suboptimal rules in decision making. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 268–283). Cambridge: Cambridge University Press.

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over- and underconfidence: the role of error in judgment processes. *Psychological Review, 101*, 519–527.

Fischhoff, B. (1982). Debiasing. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 422–444). Cambridge: Cambridge University Press.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: a Brunswikian theory of confidence. *Psychological Review, 98*, 506–528.

Harvey, N., Harries, C., & Fischer, I. (2000). Using advice and assessing its quality. *Organizational Behavior and Human Decision Processes, 81*, 252–273.

Juslin, P. (1994). The overconfidence phenomenon as a consequence of informal experimenter-guided selection of almanac items. *Organizational Behavior and Human Decision Processes, 57*, 226–246.

Juslin, P., Olsson, H., & Björkman, M. (1997). Brunswikian and Thurstonian origins of bias in probability assessment: on the interpretation of stochastic components of judgment. *Journal of Behavioral Decision Making, 10*, 189–209.

Katz, J. (1984). Why doctors don't disclose uncertainty. *The Hastings Center Report, 14*, 35–44.

Keren, G. (1997). On the calibration of probability judgments: some critical comments and alternative perspectives. *Journal of Behavioral Decision Making, 10*, 269–278.

Klayman, J. (1988). On the how and why (not) of learning from outcomes. In B. Brehmer, & C. R. B. Joyce (Eds.), *Human judgment: The SJT view* (pp. 115–162). Amsterdam: Elsevier.

Kruglanski, A. W. (1989). *Lay epistemics and human knowledge: Cognitive and motivational bases*. New York: Plenum.

Kruglanski, A. W., Webster, D. M., & Klem, A. (1993). Motivated resistance and openness to persuasion in the presence or absence of prior information. *Journal of Personality and Social Psychology, 65*, 861–876.

Lichtenstein, S., Fischhoff, B., & Phillips, L. 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.

McClelland, A. G. R., & Bolger, F. (1994). The calibration of subjective probabilities: theories and models 1980–1994. In G. Wright, & P. Ayton (Eds.), *Subjective probability* (pp. 453–482). Chichester: Wiley.

Murphy, A. H., & Brown, B. G. (1984). A comparative evaluation of objective and subjective weather forecasts in the United States. *Journal of Forecasting, 3*, 369–393.

Ronis, D. L., & Yates, J. F. (1987). Components of probability judgment accuracy: individual consistency and effects of Participant matter and assessment method. *Organizational Behavior and Human Decision Processes, 40*, 193–218.

Shanteau, J. (1988). Psychological characteristics and strategies of expert decision makers. *Acta Psychologica, 68*, 203–215.

Slovic, P., & Lichtenstein, S. (1971). Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Organizational Behavior and Human Performance, 6*, 649–744.

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

Sniezek, J. A., & Van Swol, L. M. (2001). Trust, confidence, and expertise in a judge–advisor system. *Organizational Behavior and Human Decision Processes, 84*, 288–307.

Soll, P. J. (1996). Determinants of overconfidence and miscalibration: the roles of random error and ecological structure. *Organizational Behavior and Human Decision Processes, 65*, 117–137.

Stone, E. R., & Opel, R. B. (2000). Training to improve calibration and discrimination: the effects of performance and environmental feedback. *Organizational Behavior and Human Decision Processes, 83*, 282–309.

Wells, G. L., & Murray, D. M. (1984). Eyewitness confidence. In G. L. Wells, & E. F. Loftus (Eds.), *Eyewitness testimony: Psychological perspectives* (pp. 155–170). Cambridge: Cambridge University Press.

Whitley, B. E., & Greenberg, M. S. (1986). The role of eyewitness confidence in juror perceptions of credibility. *Journal of Applied Social Psychology, 16*, 387–409.

Wright, G., & Ayton, P. (Eds.). (1994). *Subjective Probability.* Chichester: Wiley.

Wright, G., & Phillips, L. D. (1979). Personality and probabilistic thinking: an experimental study. *British Journal of Psychology, 70*, 295–303.

Yaniv, I., & Foster, D. P. (1995). Graininess of judgment under uncertainty: an accuracy–informativeness tradeoff. *Journal of Experimental Psychology: General, 124*, 424–432.

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

Yates, J. F. (1990). *Judgment and decision making*. New York: Prentice-Hall.

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

Yates, J. F., Price, P. C., Lee, J.-W., & Ramirez, J. (1996). Good probabilistic forecasters: the 'consumer's' perspective. *International Journal of Forecasting, 12*, 41–56.

Zarnoth, P., & Sniezek, J. A. (1996). The social influence of confidence in group decision making. *Journal of Experimental Social Psychology, 33*, 345–366.