Psychonomic Bulletin & Review
2002, 9 (4), 821-828

When wrong predictions provide more support than right ones

CRAIG R. M. McKENZIE and MARSHA B. AMIN
University of California, San Diego, La Jolla, California

Correct predictions of rare events are normatively more supportive of a theory or hypothesis than correct predictions of common ones. In other words, correct bold predictions provide more support than do correct timid predictions. Are lay hypothesis testers sensitive to the boldness of predictions? Results reported here show that participants were very sensitive to boldness, often finding *incorrect* bold predictions more supportive than *correct* timid ones. Participants were willing to tolerate inaccurate predictions only when predictions were bold. This finding was demonstrated in the context of competing forecasters and in the context of competing scientific theories. The results support recent views of human inference that postulate that lay hypothesis testers are sensitive to the rarity of data. Furthermore, a normative (Bayesian) account can explain the present results and provides an alternative interpretation of similar results that have been explained using a purely descriptive model.

Imagine a new geophysical theory that leads to predictions as to when large earthquakes, which are rare, will occur. Which outcome would leave you more convinced that the theory's predictions are more reliable than mere guesses, a correct prediction of a large earthquake, or a correct prediction of none? Bayesian statistics dictates that correctly predicting a rare event should be more convincing (Horwich, 1982; Howson & Urbach, 1989). Indeed, this "rarity principle" seems to be routinely exploited by researchers, who strive to correctly predict rare or surprising phenomena, presumably because such results constitute strong evidence for the theory or hypothesis being tested (but see Wallach & Wallach, 1994).

Perhaps scientists are sensitive to the rarity principle, but what about lay hypothesis testers? Some recent psychological research has indicated that people are sensitive to rarity in a qualitatively normative manner. Oaksford and Chater (1994; Oaksford, Chater, & Grainger, 1999; Oaksford, Chater, Grainger, & Larkin, 1997) have argued that participants are sensitive to rarity when performing Wason's (1968) selection task (or a related variant), and McKenzie and Mikkelsen (2000) have shown that participants considered rare confirming outcomes more supportive than common ones in a hypothesis-testing task. Nonetheless, some authors have disputed the claim that participants are sensitive to rarity when testing hypotheses. Using the selection task, Oberauer, Wilhelm, and Diaz (1999) found no convincing evidence that participants' information search strategies shifted in the direction predicted by the Bayesian account, and Evans and Over (1996) reanalyzed results from an earlier study (Pollard & Evans, 1983), which they claimed showed that participants' behavior shifted in the wrong direction. (For replies to these authors' claims, see Oaksford & Chater, 1996, in press.)

One purpose of the present article is to present new, potent evidence that lay hypothesis testers find rare outcomes highly informative. Whereas recent research has indicated that correct predictions of rare events (correct "bold" predictions) are seen as more informative than correct predictions of common ones (correct "timid" predictions; McKenzie & Mikkelsen, 2000), our first two experiments take this finding a step further: They show that even *incorrect* bold predictions are often seen as more supportive of a hypothesis than correct timid predictions.

Such findings clearly show that participants are sensitive to the rarity of data, but they raise another issue: It appears normatively suspect to deem incorrect predictions—bold or otherwise—to be more supportive than correct ones. Perhaps lay hypothesis testers sometimes apply the rarity principle when it is inappropriate to do so. However, a second purpose of this article is to show that, given some reasonable assumptions (some of which are confirmed in our third experiment), incorrect bold predictions can be *normatively* more supportive of a hypothesis than correct timid ones. Thus, our normative analysis suggests that participants might make use of rarity in a surprisingly sophisticated manner. The third and final purpose of the article is to use the normative analysis to provide an alternative account of similar results that have been explained with the use of a purely descriptive model.

This research was supported by National Science Foundation Grants SBR-9515030 and SES-0079615. The results were presented at the Annual Meeting of the Society for Judgment and Decision Making, November 2000, in New Orleans. The authors thank Karen Dobkins, Vic Ferreira, Tom Griffiths, Shlomi Sher, and Mike Ziolkowski for thoughtful comments on earlier drafts of this manuscript. Correspondence should be addressed to C. R. M. McKenzie, Department of Psychology, University of California, San Diego, La Jolla, CA 92093-0109 (e-mail: cmckenzie@ucsd.edu).

EXPERIMENT 1

Method

Participants were 105 University of California, San Diego, students who received course credit. Among other unrelated tasks in a laboratory setting, some participants read the following:

> Imagine that you have just arrived a little early for a new class on the first day of the quarter. Two other students are in the room with you. They both claim to be able to predict future events better than most people.
>
> “Okay,” you say, “make a prediction about the next person to walk into the classroom.”
>
> One student says, “The next person to walk into the classroom will be under 6 feet 8 inches tall.”
>
> The other says, “The next person to walk into the classroom will be over 6 feet 8 inches tall.”
>
> *The next person to walk into the classroom was, in fact, 6 feet 7 inches tall.*
>
> Which student do you believe is the better predictor? Choose one:
>
> ______The student who made the “under 6 feet 8 inches” prediction
>
> ______The student who made the “over 6 feet 8 inches” prediction

In this “extreme” scenario, the bold (> 6 ft 8 in.) prediction, which was wrong, was very unlikely a priori because so few people are taller than 6 ft 8 in. In contrast, the timid (< 6 ft 8 in.) prediction, which was right, was very likely a priori.

The other half of the participants were presented with a “moderate” scenario in which the only difference was that 6 ft 8 in./6 ft 7 in. was replaced with 5 ft 8 in./5 ft 7 in. Thus, (in)accuracy was held constant: One prediction was barely wrong and one was barely right. However, because about half of adults are taller than 5 ft 8 in., both predictions were moderate. The wrong prediction was no longer bold.

Half of the participants in each group were asked to explain their choices. Furthermore, the order of the predictions (and the options) was reversed for half of the participants. Among other things, this controlled for the possibility that, in the extreme scenario, the second forecaster was perceived as being strategic after hearing the first forecaster’s prediction.

Results and Discussion

The top panel in Figure 1 shows the percentage of participants who selected the forecaster who made the *wrong* prediction as the better one. The first bar corresponds to the group presented with the extreme scenario and not asked for an explanation. More than half (56%) preferred the wrong bold forecaster to the right timid one. The second bar, corresponding to participants presented with the extreme scenario and asked for a brief explanation after making their choice, shows a decrease in selecting the wrong bold forecaster, perhaps because it is easier to explain choosing the correct forecaster than it is to explain choosing the incorrect one (Tetlock, 1991). (Participants could have seen the request for an explanation before making their choice or even changed their choice after seeing the request.) Some participants might have felt that there was good reason to select the wrong bold forecaster, but were unable or unwilling to articulate the reason (see also McKenzie & Mikkelsen, 2000). Nonetheless, almost 40% preferred the wrong forecaster, and, of these 10 participants, 9 mentioned in their written explanations how rare it is to be taller than 6 ft 8 in. The right two bars show that few participants selected the wrong forecaster in the moderate scenario, regardless of whether an explanation was requested. A scenario (extreme vs. moderate) × explanation (yes vs. no) log-linear analysis revealed only an effect of scenario on the number of participants preferring the wrong forecaster [$\chi^2(1, N = 105) = 14.9, p < .001$].

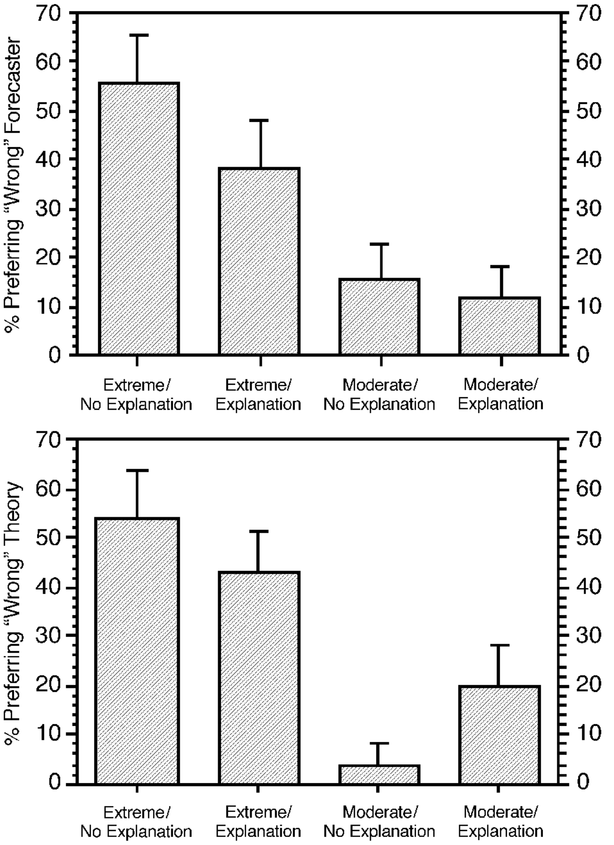

Figure 1. The top panel (Experiment 1) shows the percentage of participants who chose the forecaster who made the incorrect prediction as better than the forecaster who made the correct prediction as a function of scenario (extreme vs. moderate) and whether they were asked to provide an explanation of their choice. The bottom panel (Experiment 2) shows the results when participants chose the scientific theory they thought was most likely true. In both experiments, participants often selected the forecaster or theory that made an incorrect bold prediction over the one that made a correct timid prediction. However, when the predictions were both moderate, the forecaster or theory making the incorrect prediction was rarely chosen. Standard error bars are shown.

These results show that, holding (in)accuracy constant, the boldness of a prediction matters. A moderate forecaster who is barely wrong is rarely preferred to one who is barely right, but a bold forecaster who is barely wrong is often preferred to a timid one who is barely right.

EXPERIMENT 2

In Experiment 1, participants were asked which of two forecasters was “better”—a subjective, multidimensional judgment. In Experiment 2, participants evaluated two scientific theories and were asked which of the theories was most likely *true*. This allowed us to address whether the results of Experiment 1 were an artifact of the ambiguity of the question, which might have led to any number of interpretations. Arguably, the question posed in Experiment 2 is less open to interpretation. More generally, the complete change of context and question posed to the participants allowed for a check of the robustness of Experiment 1's findings.

Method

Participants were 113 students from the same population as in Experiment 1. Some participants read the following scenario:

> Imagine that there are two new geophysical theories that, their supporters claim, can predict when earthquakes will occur and what size they will be. The two theories are based on very different views of what causes earthquakes, and they often lead to different predictions. Because the theories are so different, it is very unlikely that both are correct.
>
> The theories are being tested in a location where small earthquakes occur every day, but large earthquakes are rare. For example, earthquakes in this location registering more than 6.0 on the Richter scale occur about once every 3 years, on average. The average daily earthquake has a magnitude of 2.0.
>
> The theories are being tested by comparing their predictions against what actually happens on a daily basis. Today was the first day of testing. Today's predictions made by the theories were:
>
> Theory A predicted that an earthquake measuring larger than 6.0 would occur. (1 out of every 1,000 earthquakes in this area is larger than 6.0.)
>
> Theory B predicted that an earthquake measuring smaller than 6.0 would occur. (999 out of every 1,000 earthquakes in this area are smaller than 6.0.)
>
> *Today, an earthquake measuring 5.9 on the Richter scale occurred.*
>
> Based on the theories' predictions about the magnitude of today's earthquake and its actual magnitude, *which theory do you think is most likely the true one*? Choose one:
>
> ______Theory A, which predicted an earthquake larger than 6.0.
>
> ______Theory B, which predicted an earthquake smaller than 6.0.

In this extreme scenario, the bold prediction (>6.0), which was wrong, was very unlikely a priori, and the timid prediction (<6.0), which was right, was very likely.

Half of the participants were presented with a moderate scenario, in which the predictions were larger/smaller than 2.0 and the earthquake measured 1.9. The participants were told that half of the earthquakes measured below 2.0 and half measured above 2.0. Thus, both predictions were moderate, but (in)accuracy was the same as in the extreme scenario: One prediction was barely right, and one was barely wrong.

As in Experiment 1, half of the participants presented with each scenario provided a brief explanation of their choice. Labeling of the theories (and the order of the options) was also reversed for half of the participants.

Results and Discussion

The bottom panel in Figure 1 shows the percentage of participants selecting the theory that made the wrong prediction as most likely to be the true theory. The left bar corresponds to participants presented with the extreme scenario and not asked for an explanation. Analogous to the results of Experiment 1, more than half of these participants (54%) preferred the theory that made the incorrect bold prediction to the theory that made the correct timid one. The second bar, corresponding to those presented with the extreme scenario and asked for an explanation, shows a slight decrease in selecting the theory that made the incorrect bold prediction. This decrease also replicates the finding in Experiment 1 and again suggests that it might be difficult to explain selecting the theory that made the wrong prediction (leading some to choose the correct timid theory). Nonetheless, 43% selected the theory that made the incorrect bold prediction, and, of these 11 participants, 8 mentioned in their explanations that earthquakes measuring greater than 6.0 are rare.

The two rightmost bars show that, when both predictions were moderate, relatively few participants selected the theory that made the wrong prediction, although more participants did so when asked for an explanation. A scenario (extreme vs. moderate) × explanation (yes vs. no) log-linear analysis on the number of participants selecting each theory revealed an effect of scenario [$\chi^2(1, N = 113) = 17.3, p < .001$]. There was also an interaction between scenario and explanation [$\chi^2(1, N = 113) = 3.9, p = .047$]: Relative to those not asked, participants asked for an explanation were less likely to select the theory making the wrong prediction in the extreme scenario and were more likely to do so in the moderate scenario. The reason for the interaction is unclear.

As in Experiment 1, these results show that the boldness of a prediction matters. This time, however, the finding occurred in the context of evaluating the truth of a theory. Despite the change in context and in the question posed to participants, a comparison of the top and bottom panels in Figure 1 reveals that the results were similar across the two experiments.

EXPERIMENT 3

Though it might seem peculiar to believe that the theory making the wrong prediction is most likely the true one, this belief can be made normatively coherent by making some simple assumptions. A general assumption required by our account is that participants treat the predictions as probabilistic. That is, participants might not interpret a prediction of an outcome (Q) as "there is a 100% chance that Q," but instead as "there is an X% chance that Q," where $X < 100$. In a separate study, we asked 75 participants how they interpreted the earthquake predictions presented in Experiment 2. Most (79%) preferred a probabilistic interpretation (e.g., "There is a *very good chance* that today's earthquake will be larger/smaller than 6.0") to a deterministic one ("Today's earthquake will *definitely* be larger/smaller than 6.0").

A further assumption is needed, however. As we will discuss in more detail later, a normative account of the results is feasible if participants consider the timid prediction to be made with more confidence than the bold one but consider the two moderate predictions to be made with about equal confidence. Although bold predictions regard rare events, this does not necessarily imply that they are perceived to be made with low confidence; bold predictions could be seen as being accompanied by at least as much confidence as timid predictions. To examine this issue empirically, in Experiment 3 we presented participants with either the extreme or the moderate scenario used in Experiment 1 (regarding the height of the next person to enter the room) and asked them which of the two predictions was made with greater confidence, or whether both predictions were made with about the same level of confidence.

Method

There were 126 participants, about half of whom were drawn from the population used in the earlier experiments, and about half of whom were paid for participating. The two scenarios were the same extreme and moderate ones used in Experiment 1, but the actual height of the next person to enter the room was eliminated. The participants simply read the introductory paragraph and the two predictions and then answered the question, "Which student do you think is more confident that his/her prediction will be correct?" There were three options, one corresponding to each of the students and a third stating, "The two students are about equally confident in their predictions." In addition to assigning participants to either the extreme or moderate scenario, we controlled for the order of the three options.

Results and Discussion

When participants were presented with the extreme scenario, the modal response (52% of the participants) was to select the "under 6 ft 8 in." forecaster as more confident, while 17% selected the "over 6 ft 8 in." forecaster, and 31% reported that both forecasters were about equally confident. When they were presented with the moderate scenario, however, the modal response (53%) was that both forecasters were about equally confident; 34% selected the "under 5 ft 8 in." forecaster as more confident, and 13% selected the "over 5 ft 8 in." forecaster. A scenario (extreme vs. moderate) × compensation (course credit vs. pay) log-linear analysis on the number of participants selecting the "under," "over," and "equal" responses revealed only an effect of scenario [$\chi^2(2, N = 126) = 6.4, p = .041$]. Those presented with the extreme scenario were more likely to select the "under" forecaster as more confident and were less likely to select the "equal" option relative to those presented with the moderate scenario.

A NORMATIVE ACCOUNT

The results of Experiment 3 make a normative account of our earlier results rather straightforward. In terms of Experiment 2, we will show that, given some reasonable assumptions, the theory making the incorrect bold prediction should be seen as more likely true than the theory making the correct timid prediction.

Uncertainty in the predictions can be represented by probability distributions. Panel A in Figure 2 shows two normal distributions, the one on the left corresponding to hypothetical confidence in the timid earthquake prediction ("smaller than 6.0"), and the one on the right corresponding to hypothetical confidence in the bold prediction ("larger than 6.0").[1] (Normal distributions are not necessary for making our points, but they are sufficient. Our goal here is not to provide a complete analysis, but to illustrate our points in a simple manner.) The two distributions have the same variance, differing only in their means. Consistent with the geophysical theories' predictions, most of the timid distribution is below 6.0 and most of the bold distribution is above 6.0, about 69% in both cases. Given these probability distributions, one can ask which theory should be believed more strongly, given an observed earthquake magnitude. The answer is the theory whose curve is highest at the observed magnitude (assuming equal prior probabilities that each theory is true, which seems reasonable in our experiments). The vertical line shows that the distribution corresponding to the timid prediction is the highest distribution at the 5.9 magnitude. Note that the two theories are equally likely to be true where the curves intersect at 6.0. Any magnitude less than that is evidence in favor of the theory making the timid prediction, and any magnitude greater than that is evidence in favor of the theory making the bold prediction.

The results of Experiment 3 show, however, that participants generally expected the timid prediction to be made with more confidence, meaning that the timid distribution will have more of its area below 6.0 than the bold distribution will have above 6.0. Panel B shows the timid distribution with about 98% of its area below 6.0, corresponding to 98% confidence in the accuracy of the prediction, while the bold distribution maintains 69% of its area above 6.0, just as in panel A. Now the bold prediction's curve is highest at the observed 5.9 value, making the theory that made the wrong prediction most likely the true one. Given the current assumptions, this signal-detection analysis is equivalent to a Bayesian one (Birnbaum, 1983; Luce, 1963; McKenzie, Wixted, Noelle, & Gyurjyan, 2001), and therefore, believing more strongly in the theory that made the incorrect prediction is Bayes optimal.

There is another factor that will likely work in favor of the theory making the bold prediction. The variances of the distributions are equal in panels A and B, but it is plausible that a bold prediction's distribution will be assumed to have less variance than that of a timid prediction. The timid prediction's distribution will probably have a wider range to cover and therefore be more diffuse. The bold prediction, occupying an extreme end of the scale, is unlikely to cover as wide a range. Panel C shows the bold distribution with decreased variance while holding constant the area above 6.0 for the bold distribution and below 6.0 for the timid distribution at 69%. As the panel illustrates, the result is that, given an earthquake magnitude of 5.9, the theory making the incorrect bold prediction is most likely the true one.

It is conceivable that bold predictions are perceived as being made with less confidence *and* as having tighter distributions. Such a case is shown in panel D. As can be seen, the theory that made the incorrect bold prediction is even more likely to be the true one under these circumstances.

In contrast to the expected asymmetry in confidence between the bold and timid predictions, recall that the

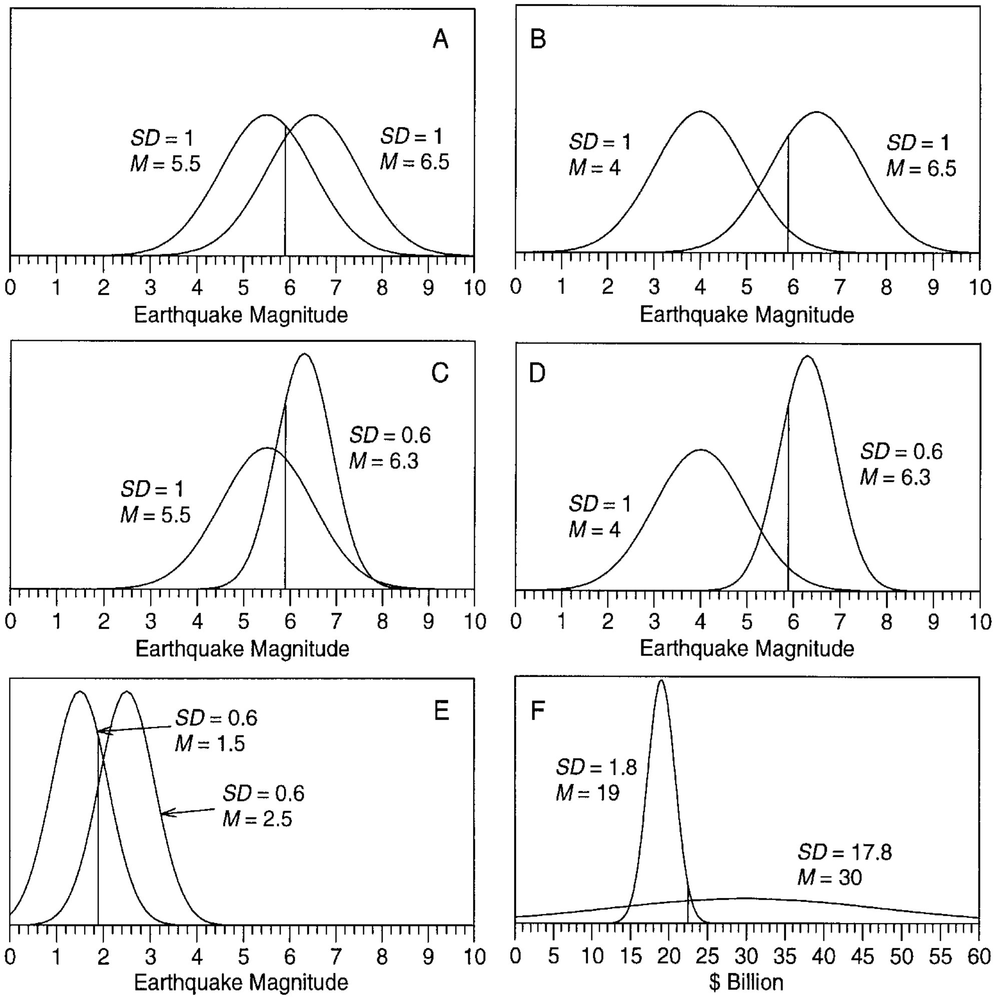

Figure 2. Panel A shows possible probability distributions corresponding to the timid prediction of earthquake magnitude ("<6.0") on the left and the bold prediction on the right (">6.0"). The two predictions are perceived to be made with the same level of confidence and intersect at 6.0. The timid distribution is therefore the highest distribution at the observed magnitude of 5.9 (vertical line), making the theory that made the correct timid prediction most likely the true theory. Panel B shows the two distributions when the bold prediction is made with lower confidence than is the timid prediction (see Experiment 3). The bold distribution has less of its area above 6.0 than the timid distribution has below 6.0, which results in their intersecting below the observed magnitude of 5.9. Thus, the curve corresponding to the bold prediction is higher at the observed magnitude, making the theory that made the incorrect prediction most likely the true theory. Panel C shows the two predictions made with equal confidence, but the bold distribution has decreased variance. This also results in the theory making the incorrect bold prediction most likely the true one. Panel D shows that when the bold prediction is made with less confidence *and* the distribution has decreased variance, the theory making the incorrect prediction is even more likely to be the true one. Panel E shows possible probability distributions for the two moderate predictions, "<2.0" and ">2.0". Because both predictions are expected to be made with the same amount of confidence (Experiment 3), the distributions intersect at 2.0, and the distribution corresponding to the theory making the correct prediction is highest at the observed value of 1.9 (vertical line). Panel F shows possible probability distributions for two judges' interval responses to a general knowledge question. Judge A's relatively flat distribution corresponds to a wide interval ($20 billion to $40 billion) and Judge B's tall distribution to a narrow interval ($18 billion to $20 billion). The vertical line represents the true value of $22.5 billion. Although the true value falls inside A's interval and outside B's interval, B's distribution is higher at the true value. Thus, B considers the true value more likely than A does.

modal participant in Experiment 3 expected the moderate predictions to be made with roughly equal confidence. In the present context, this implies that the amount of area below 2.0 for the distribution corresponding to the "smaller than 2.0" prediction should be about equal to the amount of area above 2.0 for the "larger than 2.0" prediction. Panel E illustrates two such distributions. Because of the symmetry in confidence (as in panel A), the curves intersect at 2.0, and if the observed magnitude is below 2.0, the "smaller than 2.0" theory is most likely true, and if it is above 2.0, the "greater than 2.0" theory is most likely true. Hence, the actual magnitude of 1.9, shown by the vertical line, indicates that the "smaller than 2.0" theory is most likely true.

In short, this post hoc analysis shows that both preferring the theory that made the incorrect bold prediction and preferring the theory that made the correct moderate prediction *might* be normatively coherent. We are not suggesting that our participants behaved optimally—this would depend on, among other things, the means, variances, and shapes of participants' subjective probability distributions, along with their subjective prior probabilities—only that it is plausible that their behavior makes normative sense. One key factor, empirically confirmed by Experiment 3, is that the bold prediction is perceived to be made with less confidence than the timid prediction, and that the two moderate predictions are perceived to be made with about equal confidence.

A NORMATIVE ACCOUNT OF RELATED RESULTS

The normative account also provides an alternative way to view similar results reported by Yaniv and Foster (1995), who demonstrated that precise interval estimates are sometimes viewed as superior to broad interval estimates, even when only the latter contain the true value. For example, imagine that two judges are asked how much money was spent on education by the US federal government in 1987. Judge A responds "$20 billion to $40 billion," and Judge B responds "$18 billion to $20 billion." The true value is $22.5 billion. Which judge is better? Most participants selected B, although the true value falls outside B's interval and inside A's. Yaniv and Foster (1995) asked participants many such questions and explained the pattern of preferences using a descriptive model that trades off accuracy and informativeness (where normalized error is defined as the absolute difference between the true value and the interval's midpoint, divided by the interval's width, and informativeness is defined as the log of the interval's width). The tradeoff occurs because wider intervals improve accuracy (i.e., decrease normalized error) but decrease informativeness. In the example above, though Judge A is more accurate, Judge B is more informative.

Yaniv and Foster (1995) did not consider a normative approach to the problem, but our earlier normative analysis is applicable. Panel F in Figure 2 shows possible probability distributions for the two judges' answers to the question above. Each distribution has about 42% of its area within the corresponding specified interval (e.g., 42% of the area of B's distribution is contained in the interval between $18 billion and $20 billion). The value of 42% is based in part on Yaniv and Foster (1997), who found that, when participants were asked for intervals corresponding to uncertain quantities, the intervals contained the true value between 43% and 46% of the time across three studies (see also Alpert & Raiffa, 1982; Lichtenstein, Fischhoff, & Phillips, 1982). Thus, it is reasonable to assume that participants expect others' intervals to contain the true value about this often. The vertical line corresponds to the correct value of $22.5 billion. Note that B's curve is higher than A's at that point. That is, although the true value falls outside B's interval, B is nonetheless seen as assigning a higher probability to the true value. A crucial assumption here is that the intervals reported by the judges are perceived to be relatively low confidence intervals. If, for example, participants perceive the intervals as containing the true value 98% of the time, then a higher degree of belief in the true value could not be attributed to B. As mentioned, though, participants' own intervals typically contain the true value less than 50% of the time.

We are not claiming that Yaniv and Foster's (1995) account is incorrect. Instead, our analysis makes two points. First, a normative model of Yaniv and Foster's task is feasible. This is important because it can lead to a deeper understanding of why participants behave as they do in such a task. This, in turn, can help guide the building and testing of descriptive models.

Second, it is possible that the normative account does predict behavior well. Yaniv and Foster (1995) found that their accuracy–informativeness tradeoff model outperformed several alternative models, but they did not test a normative model (probably because it was not obvious that a normative model was applicable). They also found that some alternative models performed almost as well as theirs and concluded that what all the good performing models had in common was that they traded off accuracy and informativeness. The normative model makes a similar tradeoff. Bold predictions have the advantage of tall distributions, but the disadvantage of having to be near the mark because their distributions fall away fast. Timid predictions have the advantage of maintaining some height of their curve far from the true value, but they have the disadvantage of having only modest heights even when the true value is close to their mean. In terms of the rarity principle, narrow intervals are bold predictions in that they are relatively unlikely, a priori, to contain the true value.

Although it is possible that the normative account is the best descriptive model, it is not very plausible. Not only is there much evidence indicating that people are not optimal Bayesians (e.g., McKenzie, 1994), Bayesian models are notorious for their enormous complexity even when applied to modestly complicated real-world problems (Charniak & McDermott, 1985; Dagum & Luby, 1993), making them poor candidates for models of psychological processes. Yaniv and Foster's (1995) relatively simple model, which is descriptively plausible, might capture the cognitive system's efficient solution to the complex Bayesian problem (McKenzie, 1994). Thus, the two accounts of Yaniv and Foster's (1995) results are probably best seen as complementary rather than competitive.

GENERAL DISCUSSION

The three experiments and the normative analysis make four points. First, Experiments 1 and 2 show that participants are highly sensitive to the rarity of data when testing hypotheses. Not only are correct bold predictions (correct predictions of rare events) seen as more supportive than correct timid predictions (correct predictions of common events; McKenzie & Mikkelsen, 2000; see also McKenzie, Ferreira, Mikkelsen, McDermott, & Skrable, 2001; Oaksford & Chater, 1994, 1996; Oaksford et al., 1997), but the present results show that *incorrect* bold predictions can be seen as more supportive as well. We see this as strong evidence in favor of recent views of lay inferential behavior that have postulated that people are sensitive to the rarity of data (McKenzie & Mikkelsen, 2000; Oaksford & Chater, 1994).

Second, the results of Experiments 1 and 2 are not necessarily the result of participants' applying the normative rarity principle to situations where it is inappropriate. We have shown that incorrect bold predictions are *normatively* more supportive than correct timid ones under certain conditions. An important assumption in our post hoc analysis was that bold predictions are expected to be made with less confidence than are timid predictions, which Experiment 3 confirmed empirically. This means that even if the bold prediction is off the mark, it might nonetheless arise from the theory most likely to be true, or be made by the forecaster most likely to provide the correct prediction.

Third, the normative account provides an alternative way of viewing Yaniv and Foster's (1995) finding that precise interval estimates that do not contain the true value are sometimes seen as superior to broad interval estimates that do contain the true value. Their account centered on a purely descriptive model that traded off accuracy and informativeness. The present perspective is that a judge producing a precise interval estimate that does not contain the true value might nonetheless be more likely to produce the correct answer. A key assumption here is that the reported intervals are seen as relatively low-level confidence intervals (e.g., of around 50% rather than 98%), and this appears reasonable given that participants' own intervals contain the true value less than 50% of the time (Alpert & Raiffa, 1982; Lichtenstein et al., 1982; Yaniv & Foster, 1997). A disadvantage of the normative account is that assumptions have to be made about the underlying probability distributions. It is possible that Yaniv and Foster's (1995) model, which is simpler and psychologically more plausible than the normative Bayesian account, captures the cognitive system's efficient solution to the complex normative problem. Regardless of its descriptive status, the normative analysis deepens our understanding of the task and of why participants behave as they do.

Finally, these results suggest that scientists and forecasters are even better off making bold predictions than they perhaps realize: Making a bold prediction is less risky than it appears because it can be wrong and still be convincing. People are more willing, for good reason, to tolerate inaccuracy when a prediction is bold, which can lead an otherwise disconfirmatory outcome to be perceived as confirmatory.

REFERENCES

Alpert, M., & Raiffa, H. (1982). A progress report on the training of probability assessors. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 294-305). New York: Cambridge University Press.

Birnbaum, M. H. (1983). Base rates in Bayesian inference: Signal detection analysis of the cab problem. *American Journal of Psychology*, **96**, 85-94.

Charniak, E., & McDermott, D. (1985). *An introduction to artificial intelligence.* Reading, MA: Addison-Wesley.

Dagum, P., & Luby, M. (1993). Approximating probabilistic inference in Bayesian belief networks is NP-hard. *Artificial Intelligence*, **60**, 141-153.

Evans, J. St. B. T., & Over, D. E. (1996). Rationality in the selection task: Epistemic utility versus uncertainty reduction. *Psychological Review*, **103**, 356-363.

Horwich, P. (1982). *Probability and evidence*. Cambridge: Cambridge University Press.

Howson, C., & Urbach, P. (1989). *Scientific reasoning: The Bayesian approach.* La Salle, IL: Open Court.

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

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology: I* (pp. 103-189). New York: Wiley.

McKenzie, C. R. M. (1994). The accuracy of intuitive judgment strategies: Covariation assessment and Bayesian inference. *Cognitive Psychology*, **26**, 209-239.

McKenzie, C. R. M., Ferreira, V. S., Mikkelsen, L. A., McDermott, K. J., & Skrable, R. P. (2001). Do conditional hypotheses target rare events? *Organizational Behavior & Human Decision Processes*, **85**, 291-309.

McKenzie, C. R. M., & Mikkelsen, L. A. (2000). The psychological side of Hempel's paradox of confirmation. *Psychonomic Bulletin & Review*, **7**, 360-366.

McKenzie, C. R. M., Wixted, J. T., Noelle, D. C., & Gyurjyan, G. (2001). Relation between confidence in yes–no and forced-choice tasks. *Journal of Experimental Psychology: General*, **130**, 140-155.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, **101**, 608-631.

Oaksford, M., & Chater, N. (1996). Rational explanation of the selection task. *Psychological Review*, **103**, 381-391.

Oaksford, M., & Chater, N. (in press). Optimal data selection: Revision, review, and reevaluation. *Psychonomic Bulletin & Review*.

Oaksford, M., Chater, N., & Grainger, B. (1999). Probabilistic effects in data selection. *Thinking & Reasoning*, **5**, 193-243.

Oaksford, M., Chater, N., Grainger, B., & Larkin, J. (1997). Optimal data selection in the reduced array selection task (RAST). *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **23**, 441-458.

Oberauer, K., Wilhelm, O., & Diaz, R. R. (1999). Bayesian rationality for the Wason selection task? A test of optimal data selection theory. *Thinking & Reasoning*, **5**, 115-144.

Pollard, P., & Evans, J. St. B. T. (1983). The effect of experimentally contrived experience on reasoning performance. *Psychological Research*, **45**, 287-301.

Tetlock, P. E. (1991). An alternative metaphor in the study of judgment and choice: People as politicians. *Theory & Psychology*, **4**, 451-475.

Wallach, L., & Wallach, M. A. (1994). Gergen versus the mainstream: Are hypotheses in social psychology subject to empirical test? *Journal of Personality & Social Psychology*, **67**, 233-242.

Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, **20**, 273-281.

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

Yaniv, I., & Foster, D. P. (1997). Precision and accuracy of judgmental estimation. *Journal of Behavioral Decision Making*, **10**, 21-32.

NOTE

1. In order to be consistent with the reported experiments, we are interpreting the probability distributions in terms of a person's subjective confidence over the various earthquake magnitudes for each theory's prediction. Alternatively, however, the distributions could be interpreted as likelihood distributions of earthquake magnitudes actually predicted by each theory.

(Manuscript received October 18, 2000;
revision accepted for publication December 21, 2001.)