Journal of Experimental Psychology:
Learning, Memory, and Cognition
1989, Vol. 15, No. 4, 605-619

0278-7393/89/$00.75

Outcome Feedback: Hindsight *and* Information

Stephen J. Hoch and George F. Loewenstein
University of Chicago
Center for Decision Research
Graduate School of Business

Although "hindsight bias" research has demonstrated that outcome feedback leads people to exaggerate the odds they would have placed on known outcomes, learning theories view feedback as the key to effective adaptation. Our model and data show that outcome feedback has multiple effects. People can extract diagnostic information from feedback despite overestimating what they would have known in foresight. Ss reliably discriminate easy tasks where they "knew it all along" from difficult tasks where they "never would have known it"; differential reactions to feedback provide information useful in other judgment tasks such as assessment of population base-rates and personal knowledge calibration. In Experiments 1–4, feedback increased judgmental accuracy by over 150%. However, a final experiment suggests that certain tasks (insight problems) produce such strong "I knew it all along" reactions that hindsight can overwhelm the information contained in feedback and reduce predictive accuracy.

On a daily basis, we are bombarded by information about outcomes. The stock market plummeted, a nuclear reactor partially melted down, a friend returned safely from Africa, we ran out of gas. In processing this information, it is often useful to recapture what we expected to occur before receiving the information. Accurate memory of prior expectations helps us to respond constructively to outcome feedback. Discrepancies between expectations and outcomes may indicate a deficiency in our mental model of the environment. Events that jibe with expectations should reinforce our current perspective. Hence, exaggerating our judgmental prowess by remembering our expectations as having been more accurate than they really were can lead to overconfidence, preventing warranted updating of our mental model, whereas underestimating the accuracy of prior expectations can result in unjustified revisions. After an uncertain outcome has occurred, it is important to discriminate between those situations where we "knew it all along," those where we "never would have known it," and others that fall somewhere in between.

A closely related skill, applying to situations where we already possess outcome information, involves imagining how the current situation would have appeared to us and what we would have expected to occur if we had not received outcome information. This skill is particularly useful for assessing the opinions or knowledge of other people who possess less information than we do. For example, in plea bargaining, the accused, possessing full knowledge of his or her own culpability, must decide whether to accept the prosecutor's offer on the basis of an assessment of the evidence he or she possesses. In preparing a convincing lecture or paper, it is important to accurately anticipate the ability and motivation of the audience to process material with which we are overly familiar.

During the past two decades an expanding body of research has examined people's ability, given outcome feedback, to adopt the prefeedback perspective. The main focus of this research has been on the "hindsight bias," a term coined by Fischhoff (1975), referring to the tendency for people to "consistently exaggerate what could have been anticipated in foresight" (Fischhoff, 1982).

A Brief History of Hindsight

The existence of hindsight bias has been supported by a set of convincing empirical investigations using a wide range of predictive tasks: real political developments, obscure historical events, medical diagnoses, and answers to trivia questions. These studies have employed two basic paradigms.

In within-subjects studies, subjects attempt to recall the beliefs they held in the past before receiving outcome feedback. For example, in one study (Fischhoff & Beyth, 1975) subjects judged the probability of different outcomes of Nixon's then forthcoming trips to Moscow and Peking. After the trips they were told how these outcomes were resolved and were asked to recall the probabilities they had earlier assigned to them. Remembered probabilities were slanted in the direction of actual outcomes; subjects remembered having assigned higher probabilities to events that actually occurred and lower probabilities to events that did not occur.

Between-subjects designs provide subjects with information concerning an uncertain event. On the basis of this information alone, one group of subjects is asked to estimate the probability that the event will occur. Another group receives outcome feedback and is asked to estimate the likelihood that they would have assigned to the event if they had not been informed of the outcome. Hindsight is then assessed by comparing the probability judgments of the two groups. The

This research was supported in part by the Bozell, Jacobs, Kenyon, and Eckhardt Faculty Endowment Fund, the IBM Faculty Research Fund at the Graduate School of Business, and the Alfred P. Sloan Foundation.

Correspondence concerning this article should be addressed to Stephen J. Hoch, University of Chicago, Graduate School of Business, 1101 E. 58th Street, Chicago, Illinois 60637.

typical finding is that subjects who receive outcome feedback assign higher probabilities to events that occur and lower probabilities to events that do not occur than subjects who do not receive feedback (Fischhoff, 1977; Hasher, Attig, & Alba, 1981). In a variant of this design experimental subjects are asked not to second guess their own judgments, but to guess the probability that subjects not informed of the outcome would confer on the target event (Fischhoff, 1975; Wood, 1978). These two designs provide comparable results, suggesting that subjects asked to predict judgments of their peers attempt to assess the probability that they would have assigned in the absence of feedback and then to adjust from their own level of uncertainty to predict that of other people—a process we term *projection* (Hoch, 1987).

The overwhelming verdict from this work is that hindsight bias is a robust phenomenon that is not easily eliminated or even moderated (Fischhoff, 1982). Fischhoff (1975) found that hindsight remained when subjects were explicitly told to ignore the outcome knowledge. Warnings to pay close attention or being told that people typically exaggerate their knowledge did not reduce hindsight (Fischhoff, 1977). And, although hindsight was reduced in the within-subjects designs mentioned earlier, this reduction was in large measure due to the fact that subjects were able to remember their exact answers about two-thirds of the time (Fischhoff, 1977; Wood, 1978). Professional education and attendant expertise also had no effect on the magnitude of hindsight bias (Arkes, Wortmann, Saville, & Harkness, 1981; Mitchell & Kalb, 1981). However, generating counterfactual explanations for why the outcome could have turned out differently did reduce the magnitude of the bias (Slovic and Fischhoff, 1977; also see Hoch, 1985; Koriat, Lichtenstein, & Fischhoff, 1980). Camerer, Loewenstein, and Weber (in press) found that monetary incentives and feedback alone did not reduce the bias, but market transactions involving interactions between traders who exhibited different levels of hindsight bias reduced the overall bias by one half. And, although Hasher et al. (1981) managed to eliminate the bias, they had to go to extreme lengths by discrediting the outcome feedback in such a way that subjects realized that it was completely unreliable.

The Paradox of Outcome Feedback

The hindsight bias, when first identified, flew against the conventional wisdom that outcome information would help or, at worst, not adversely affect predictive accuracy.

> If we know what has happened and what problem an individual was trying to solve, we should be in a position to exploit the wisdom of our own hindsight in explaining and evaluating his or her behavior. Upon closer examination, however, the advantages of knowing how things turned out may be oversold (Fischhoff, 1975). In hindsight, people consistently exaggerate what could have been anticipated in foresight. They not only tend to view what has happened as having been inevitable but also to view it as having appeared "relatively inevitable" before it happened. People believe that others should have been able to anticipate events much better than was actually the case. They even misremember their own predictions so as to exaggerate in hindsight what they knew in foresight (Fischhoff & Beyth, 1975). (Fischhoff, 1982, p. 341)

While the research on hindsight has focused on the distortional effects of outcome feedback, the learning literature has viewed feedback as the key to effective adaptation (Anderson, 1983; Brunswik, 1952; Einhorn & Hogarth, 1978; Hogarth, 1981; Nelson, 1971). How can these seemingly divergent views of feedback be reconciled? We argue that outcome feedback has multiple effects, only one of which involves retrospective increases in confidence that one would have anticipated what actually did happen. Although hindsight researchers have not denied the potential usefulness of outcome feedback (e.g., Fischhoff 1975), studies of hindsight have, perhaps inadvertently, distracted attention from other, beneficial effects of outcome feedback. We present data demonstrating that even when people process feedback in a biased hindsightful manner, the same feedback can simultaneously provide diagnostic information useful in other judgment tasks such as the assessment of population base-rates and personal knowledge calibration. After laying out a model detailing the multiple effects of outcome feedback, we discuss cognitive mechanisms that might explain how feedback, even when distorted by hindsight, can still be informative.

Multiple Effects of Outcome Feedback

Previous hindsight studies have shown that subjects realign their retrospective judgments in line with actual outcomes; they tend to exaggerate the accuracy of prior judgments. This has the effect of reducing the amount of feedback they perceive as disconfirming. Yet even when distorted by hindsight, feedback may improve judgmental accuracy. Consider an overconfident person who is actually correct 60% of the time but believes she is correct 80% of the time. How will feedback affect her overconfidence? Although the individual should ideally receive negative feedback 40% of the time, hindsight will reduce feedback perceived as disconfirming to below this rate. If hindsight is severe, negative feedback could be reduced to below 20% and would, in fact, exacerbate overconfidence. If hindsight is less pronounced, however, the rate of disconfirming feedback could fall between 20% and 40%, leading to a reduction in overconfidence even though feedback is distorted by hindsight.

Hindsight is not an all-or-none phenomenon. Rather than always eliciting an "I knew it all along" reaction, outcome feedback may produce a different response—more along the lines of "I never would have known it"—when subjects either (a) have no familiarity with the outcome or (b) start off being confident in what turns out to be the wrong outcome. We examine whether these differential reactions provide diagnostic information helpful to subjects in discriminating between easy and difficult prediction tasks and in judging the probability that they and others could and would have made a correct prediction. To illustrate, consider hypothetical reactions to outcome feedback about two questions, one quite easy and one quite difficult.

Given the question "What is the largest desert on earth?", one might naturally respond "the Sahara," thinking back to Rommel in World War II and the publicity surrounding the droughts and famines in North Africa in more recent times. The question appears to be easy (and in truth is easy—92%

of our subjects answered correctly), but doubt may creep in (the average level of confidence for this question was only .67). What about the Gobi desert in central Asia? It is not as well known, but the Gobi is big and so is Asia. And is it not true that almost all of Western Australia is desert? A consistent finding in the probability calibration literature is that although people are typically overconfident about their level of knowledge, they show systematic underconfidence on easy items (Lichtenstein & Fischhoff, 1977). In this situation, learning that the correct answer is "the Sahara" gives you important information; the question was indeed easy, because the correct answer is the response that first came to mind. For relatively easy predictions, the provision of outcome feedback clearly leads to the ubiquitous "I knew it all along" reaction; but for easy predictions, this reaction is usually quite justified and may lessen underconfidence.

The effect of outcome feedback on difficult problems is more complicated. Here it is important to distinguish between two types of response errors (Krinsky & Nelson, 1985). Problems can be difficult either because subjects make errors of commission (by giving the wrong answer) or errors of omission (by not generating any response). Feedback is likely to be especially informative for problems prone to errors of commission, where overconfidence is typically extreme (Lichtenstein & Fischhoff, 1977).

Consider the question: "Mount Kilimanjaro is located in which country?" Suppose that you do not know the answer for sure, but you have some familiarity with the domain (you have heard of Kilimanjaro) and are reminded of Hemingway, safaris, and Africa, so you guess Kenya. Despite the haziness of your information, if you are a typical subject, you would be overconfident in your response (Lichtenstein, Fischhoff, & Phillips, 1982). Hence, when informed that the answer is actually Tanzania, you may experience a certain sense of surprise because your expectations have been violated. (Subjects indicated an average probability rating of .57 in the correctness of their answer despite the fact that only 3% produced the correct response.) Whereas control subjects are overconfident in their erroneous belief that the item is relatively easy, the surprise experienced by the feedback group is diagnostic, indicating that the question is probably difficult, both for themselves and others. Even though this "never would have known it" reaction may be muted by hindsight bias (e.g., an increase in retrospective probability due to a rationale like "It's tough, but there's a chance I might have gotten lucky"), nevertheless the surprise accompanying feedback acts as an antidote to the normal overconfidence.

For errors of omission, subjects cannot be overconfident about their performance because they were not able to generate an answer and will recognize that they have no chance of being correct. Nevertheless, feedback may still be informative for related judgments. For example, even though subjects are unable to generate a response, they might be overly confident that they would be able to if given more time; or they might overestimate the fraction of their peers who would answer correctly. Feedback could reduce overconfidence for such related judgments. Alternatively, a complete retrieval failure sometimes may make subjects overly pessimistic; upon feedback, subjects may recognize that the item is less difficult than originally believed.

A Model of Hindsight and Information

Figure 1 illustrates how feedback can lead to hindsight and also provide information that leads to improved performance on other judgment tasks. Consider a group of people (control) asked to first answer a series of two alternative, multiple-choice items of varying difficulty and then to indicate the probability that their response (R) is the correct answer (A)—that is, $P(R = A)$. Another group of people (feedback) are informed of the correct answer to each question and then asked to indicate which response (Alternative 1 or 2) they *would have* chosen and the probability they *would have* assigned to their response being correct. We can order the items from easy to difficult on the basis of the frequency of correct responses. In the top diagram we plot the level of confidence that subjects assign to the *correct* response [$P(A_c)$] against the actual base rate, a standard operationalization of hindsight. Because only the feedback group knows with certainty the correct answers at the time of judgment, $P(A_c)$ is calculated indirectly from the raw certainty ratings. For correct responses ($R=A$), confidence in the correct outcome is $P(A_c)=1-P(R=A)$;

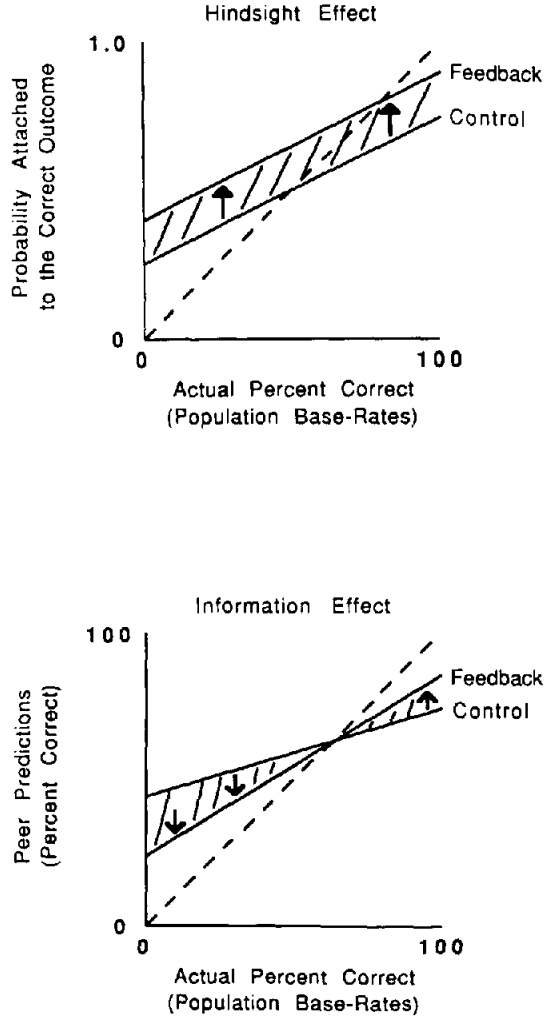

Figure 1. Hypothetical data patterns illustrating feedback-induced hindsight effects and information effects.

for incorrect responses ($R \neq A$), confidence in the correct outcome is $P(A_c) = 1 - P(R = A)$. The diagonal represents the ideal where subjects, on average, assign probabilities to the correct alternative matching the underlying base-rates.

The "control" line represents the pattern these judgments would take if subjects were unable to discriminate perfectly between easy and difficult items; subjects' responses are regressive. Feedback-induced hindsight will cause an ex post exaggeration of the likelihood of anticipating what is in fact the correct response. Hence the line marked "feedback" is simply the control line shifted upward. The hindsight effect is captured by the hashed area between the feedback and control lines. In the top half of Figure 1, we have shown hindsight as a simple intercept shift, where feedback subjects exaggerate $P(A_c)$ at a constant rate across differing base rates. In one study, however, Fischhoff (1977) found that hindsight was more pronounced on difficult than on easy items. Such an effect would reduce the slope of the feedback line, and the overall hindsight effect (the hashed area) would result from an increase in intercept and a decrease in slope.

Even in the presence of the hindsight displayed in the top diagram, however, outcome feedback may provide information (unavailable to the control group) which improves the ability to discriminate between easy and difficult items. To see this, consider how outcome feedback influences performance on a different but related task. In the lower diagram of Figure 1, the vertical axis now represents subjects' predictions concerning the difficulty of each item, that is, base-rate predictions about the percentage of peers who would answer an item correctly. In making such peer predictions, previous research has shown that subjects rely heavily on how much they think they personally know as a basis for their projections about their peers (Nickerson, Baddeley, & Freeman, 1987; Travers, 1943). This task differs in a subtle but important way from the task used to demonstrate hindsight. In the hindsight task, we are interested in the probability that subjects assign to the correct outcome, $P(A_c)$. In the peer prediction task, the focus is on the probability of being correct, irrespective of which outcome actually occurs.[1] While the feedback group always knows the correct answers to the questions and must make a subjective assessment of whether they would have been correct in foresight, control subjects do not know the correct answers with certainty and have to rely on what they believe to be the correct answer. What this means is that control and feedback subjects sometimes are not rating the same thing when making peer predictions. On easy items, control and feedback subjects typically are rating the same answer; however, on difficult items, control subjects may be rating a wrong answer that they think is correct (and so are bound to be overconfident) while feedback subjects are focusing on the ability of their peers to specify the correct answer.

Now how does all of this map onto Figure 1? Again, control group predictions are regressive because of the inability to perfectly discriminate easy from difficult items. In contrast, let us assume that on the receipt of feedback, subjects experience differential reactions depending on level of difficulty—that is, "I knew it all along" reactions on easy and "I never would have know it" reactions on difficult items. If subjects use their personal reactions to feedback as a basis for their peer predictions, then we might expect the line representing feedback subjects' responses to pivot more into line with the objective likelihoods, signifying improved calibration of base-rate predictions.

As an example, go back to the Mt. Kilimanjaro question. If feedback subjects do experience a "I never would have known it" reaction to Tanzania, they may realize that the item is difficult (only 18% correct on the two-alternative, multiple-choice version) and that it is unlikely that they or their peers would have answered correctly. Even if feedback subjects display hindsight by exaggerating the likelihood of choosing Tanzania, it is unlikely (at least according to our data) that they would be as overconfident as control subjects. Control subjects typically believe quite confidently that the answer is Kenya and are rating the percent of peers who would choose (incorrectly) the Kenya alternative. If our analysis is correct, one way that the information effects of feedback arise is that the control group is "blissfully" ignorant of their overconfidence while the feedback group has experienced some surprise which partially moderates that overconfidence. And as mentioned earlier, the information effect may also arise by correcting underconfidence on easy items. Hindsight, on the other hand, arises not because subjects deny that they have learned anything from feedback, but rather because they consistently underestimate exactly how much they *have* learned from feedback; for example, instead of experiencing the "I never would have known it" reaction, they experience a less extreme "I *probably* would not have known it" reaction. Hindsight may be a by-product of a learning process where subjects imperfectly match feedback to existing knowledge—information is the upside and hindsight is the downside.

Cognitive Mechanisms Underlying Hindsight and Information

Assessing item difficulty after receiving feedback requires people to assess the likelihood that they *would* have been able to recall an item that they now know. Our analysis of information effects depends crucially on how faithfully people can simulate the act of retrieval. The cognitive task is analogous to that entailed in assessing "feeling-of-knowing" (Blake, 1973; Hart, 1965; Nelson, Gerler, & Narens, 1984). In the current case, subjects need to assess difficulty of recall for items where recall has been preempted (because they already know them through feedback); to assess feeling-of-knowing, subjects must evaluate the familiarity of items they cannot actually recall. Research indicates that feeling-of-knowing is, in fact, diagnostic. Subjects are able to discriminate reliably between chronic (persistent) and only temporary (momentary forgetting) memory failures (Nelson, Leonesio, Landwehr, & Narens, 1986). Feeling-of-knowing judgments have been shown to be good predictors of performance in many cogni-

[1] These percent-correct peer predictions are not equivalent to the peer predictions used by Fischhoff (1977) and Wood (1978). They used a peer prediction task to assess hindsight by examining the ratings that their subjects said a group of peers would have assigned to the alternative that turned out to be the correct answer [$P(A_c)$].

tive tasks including recognition, recall, perceptual identification, and relearning (e.g., Nelson et al., 1984).

Two mechanisms have been adduced to explain the accuracy of feeling-of-knowing ratings: "trace-access mechanisms" and "inferential mechanisms" (Nelson et al. 1984; Yaniv & Meyer, 1987). Trace-access refers to a true "tip of the tongue" experience (Brown & McNeill, 1966; Koriat & Leiblich, 1974) where people have partial access to a memory trace of a target item. Inferential mechanisms are employed when subjects, by assessing their related knowledge of a subject area, infer that there is a high likelihood they have been exposed to the unrecalled item. Similar mechanisms also may be responsible for peoples' ability to derive diagnostic information from feedback.

Analogous to "tip of the tongue," subjects may experience a "rings a bell" sensation on receiving outcome feedback. Such a response indicates that the true answer was indeed available in memory; its strength provides a clue concerning just how likely the subject would have been to recall the information. Alternatively, subjects may experience a total lack of familiarity with the feedback, suggesting that they would not have been able to generate the correct answer. Feedback can also aid in item discrimination through inference-based processes. Congruence of outcome feedback with existing knowledge provides a clue that the item was easy. Outcome feedback at variance with item-relevant knowledge signals a difficult problem. For example, the feedback that Kilimanjaro is in the Alps would most likely be highly incongruent with most people's prior expectations and related geographical knowledge. Hence, a subject who received credible feedback to that effect would most likely judge the question to be extremely difficult.

The distinction between trace-access and inferential mechanisms is related to two-phase models of recognition memory (Atkinson & Juola, 1974; Gillund & Shiffrin, 1984; Juola, Fischler, Wood, & Atkinson, 1971; Mandler, 1972, 1980). These two-phase models postulate that recognition performance is based on a quick reading of item familiarity (analogous to trace access) except when familiarity is low (below some threshold), at which point subjects must engage in a more extended memory search (inference-based mechanisms) to determine whether the item is old or new (Gentner & Collins, 1981; Glucksberg & McCloskey, 1981; Reder, 1982). Although both trace-access and inferential mechanisms can result in useful information from outcome feedback, both contain potential pitfalls. Familiarity (recognition) and likelihood of recall are related, but the former provides only an imperfect signal of the latter. Indeed a failure to adequately distinguish between recognition and recall may, in part, underlie the hindsight bias. Subjects, *recognizing* an outcome that occurs, may be overconfident that they actually would have *generated* that particular outcome.

Some of the mechanisms hypothesized to play a role in hindsight can interfere with the two mechanisms discussed above. Fischhoff used the term "creeping determinism" and Hasher et al., (1981) "erase-update" to capture the idea that subjects assimilate outcome information into existing knowledge structures, basically wiping out alternative outcomes and causal sequences. If outcome feedback has such an effect, it could interfere with the diagnosticity of trace-access responses. If, on reception of outcome information, the new information is stored in memory and existing incongruent knowledge erased, then subjects may have "rings a bell" responses to items that are, in fact, entirely novel. However, Hasher et al. (1981) found conditions where subjects could retrieve their prefeedback perspectives, casting doubt on strong versions of the memory-change explanations. In related research Gardiner and Klee (1976; Klee & Gardiner, 1976) also found that subjects were able to accurately assess their prefeedback state of knowledge. Their subjects first engaged in a standard free-recall task and later were re-presented with the original word list; subjects were quite accurate in their ability to discriminate those items that they had successfully recalled on the previous test from those that they had not recalled. If we think of the re-presented list as a form of "outcome feedback," then these results may indicate that subjects have some ability to separate those things they knew from those that they did not know.

Other explanations for hindsight could interfere with the inference-based mechanisms. These include selective recall of outcome-consistent information (Dellarosa & Bourne, 1984) and reduced motivation to engage in comprehensive search (Hoch, 1985) when reasoning backwards from effect to cause (Einhorn & Hogarth, 1981). Reduced search could also entail diminished reasoning about the congruence of outcome feedback with existing knowledge. Outcome-influenced recall could occlude incongruent information and focus attention on information that is congruent with the outcome feedback. All of these phenomena would reinforce overconfidence on difficult items, undermining peoples' ability to derive useful information from feedback.

Summary

In sum, it is possible, but by no means obvious, that outcome feedback will help subjects to discriminate between easy and difficult items. Whether it does depends on the degree to which subjects can accurately assess familiarity and can objectively infer the congruence of outcome feedback with existing relevant knowledge. At one extreme it is possible that subjects could extract information from their own reactions to feedback without experiencing any hindsight, though such a finding seems unlikely, given the robustness of the hindsight effect. At the other extreme, subjects could experience hindsight sufficiently powerful to wipe out the potentially beneficial effects of feedback. If subjects *always* experience strong "I knew it all along" reactions, even on difficult items, then feedback will not aid in item discrimination. In terms of Figure 1, this would be represented by a total flattening out of the feedback line in the upper figure. However, if subjects do have some ability to retrospectively simulate memory processes and assess the congruence of outcome feedback with existing knowledge, then outcome feedback can improve future judgmental accuracy even in the presence of some hindsight at all levels of difficulty.

Five experiments were conducted to examine how feedback-induced hindsight and information combine to influence predictive accuracy. The first study, using a standard forced-choice recognition paradigm, was designed to measure

both the level of hindsight bias and the effect of feedback information on predictive accuracy and discriminatory power. Studies 2 and 3 were conducted to generalize the joint effects of hindsight and information, utilizing a general knowledge recall task. In Study 4 we directly measured feedback-induced surprise by using both recognition and recall procedures. Study 5 examined the influence of feedback about insight and noninsight tasks requiring extended deliberation.

Experiment 1

Method

The first study examined the effects of outcome feedback both on retrospective judgments of previous knowledge levels and on the accuracy of prospective judgments involving the same stimuli. Even if hindsight is a natural by-product of the outcome feedback process, a legitimate question is whether and how people capitalize on the information contained in feedback in later tasks and situations. For example, how does outcome feedback influence prospective predictions about other people's levels of knowledge? And how does outcome feedback influence personal knowledge calibration?

Procedure. The experimental task was a variant of what is known as a "half-range" task used in numerous studies of knowledge calibration (see Lichtenstein et al., 1982). Subjects were presented with statements accompanied by two alternatives. For example, "The predominant religion in Albania is (a) Eastern Orthodox or (b) Moslem." In the no-feedback control condition, subjects first had to select which of the alternatives they believed correctly completed the statement by circling either *a* or *b*. Next, subjects had to indicate their level of confidence in their selected answer. Subjects indicated their own level of certainty (designated as *o*) using a 6-point probability scale, ranging from 0.5 (*absolutely uncertain about answer*) to 1.0 (*absolutely certain answer is correct*). In addition, subjects were told that a 0.5 probability rating "indicates that you believe it is just as likely that your answer is correct as incorrect." The half-range probability scale is appropriate because subjects who feel less certain than 0.5 in the chosen answer should choose the alternative response.

After completing these two tasks for each of 20 general knowledge questions, subjects moved on to the next task. This required subjects to make predictions about the performance of a target population, their graduate student peers, on the same 20 general knowledge questions. Target norms (*t*) were determined by the performance of control subjects from all experiments reported herein. The subject's task was to predict the percentage of peers who selected the correct alternative to each of the questions. These predictions (designated as *p*) were made using a 0% to 100% scale.

Subjects in the outcome feedback condition first were shown the 20 statements, with the correct alternative underlined. They were told to study each item carefully because they would be asked questions about them in later stages of the experiment. Next, for each item they were asked to indicate which of the alternatives they "*would have* believed to be correct," assuming that they had not just seen the correct answer, by circling *a* or *b*. After selecting an alternative, they were asked to indicate how certain (*o*) they "*would have* been that your choice was correct" (the alternative just circled) by using the same 6-point half-range scale as that used by the control subjects. Finally, like the control subjects, they were asked to predict the percentage of their peers (*p*) who selected the correct alternative for each question.

Stimuli. The stimuli were trivia-like, general knowledge questions gleaned from a variety of sources, with most of the items coming from the stimuli used by Lichtenstein and Fischhoff (1977) in early studies of personal knowledge calibration. The questions covered a broad spectrum of subject areas including history, religion, medicine, business, geography, and so forth. In order to ensure that the questions varied in difficulty [*P*(correct)], 40 questions were given to a group of graduate students in the subject population. From this larger set of items 20 questions were selected, ranging from very difficult ("Aladdin's nationality was [a] Persian or [b] Chinese" [12% correct]), to moderately difficult ("The number of fluid ounces in a pint is [a] 8 or [b] 16" [70% correct]), to quite easy ("The Greek god of wine is [a] Apollo or [b] Bacchus" [91% correct]). The difficult questions consisted both of items that were commission-like (e.g., Aladdin) and items that were more omission-like in that subjects have little idea which answer is correct ("The Galapagos Islands belong to [a] Ecuador or [b] Peru," [44% correct]). Each of the questions was listed in the same order on both the control and feedback versions of the questionnaire. The *a* and *b* alternatives each were correct 50% of the time.

The variation in question difficulty was important for three reasons. First, it would facilitate our analysis of the accuracy of peer predictions. Second, Fischhoff (1977) previously found that hindsight bias was more pronounced for hard than easy items. We were interested in how, if at all, this differential hindsight would influence the information value of outcome feedback. If hindsight bias were sufficiently greater on difficult items (i.e., not just an upward shift as in Figure 1, but an upward shift coupled with a flattening out of slope), then it is possible that hindsight could overwhelm any information effect. Finally, and most important, the hypothesized beneficial effect of outcome feedback on discrimination of easy and difficult items depends crucially on the existence of variation in item difficulty.

Subjects and design. Subjects were 108 MBA students at the Graduate School of Business at the University of Chicago. The experiment was administered to two introductory business classes in the first week of the course. The only independent variable involved the presence or absence of outcome feedback. Subjects were randomly assigned to the two feedback conditions. All stimuli and instructions were contained in experimental booklets. The task was self-paced, requiring approximately 20 min.

Results

A variety of dependent variables were analyzed in terms of both feedback condition and actual item difficulty. The analyses took two forms. For some of the analyses, the 20 questions were divided into three levels of difficulty based on the target (*t*) norms: 7 difficult items (28% correct on average), 6 medium items (70% correct), and 7 easy items (88% correct). Individual subject-level regression and correlation analyses were also conducted.[2] As will become apparent from the regression analyses, the results were insensitive to the exact splits used to control for item difficulty. In all data analysis here and in the experiments to follow, the own-certainty probability ratings were rescaled to a 0–100 percentage scale ($o \times 100$) to render them comparable to the peer predictions (*p*) and target norms (*t*).

[2] Two types of correlation coefficients were computed: a parametric Pearson product-moment correlation and an ordinal Goodman-Kruskal gamma. Gammas are reported in the text, but for all experiments the two measures of association were very similar and always led to the same statistical inference. The only difference was that the gammas tended to be slightly smaller in overall magnitude. In all significance tests involving Pearson correlations, the individual coefficients were first transformed by using Fisher's *r* to *z* transformation.

Hindsight analyses. The hindsight effect was operationalized as the difference between the probability rating that feedback and control subjects assigned to the correct alternative. When subjects selected the correct alternative (e.g., they circled *a* when *a* was correct), then their responses were simply the probability rating they used to indicate their level of certainty (i.e., *o*). However, when they (both control and feedback subjects) selected the incorrect alternative, their probability ratings were transformed into certainty ratings for the correct alternative by using the transformation $(100 - o)$. For instance, a subject who chose the incorrect alternative and assigned a level of certainty of 80 would have his or her probability rating transformed from 80 for the incorrect alternative to 20 for the correct alternative $(100 - 80 = 20)$. Hindsight is evident whenever the feedback group assigns higher probability ratings to the correct responses than does the control group (also see Campbell & Tesser, 1983).

Repeated measures multivariate analyses of variance (MANOVAs) indicated that feedback subjects ($M = 69$) assigned higher probabilities to the correct answer than did controls ($M = 62$), $F(1, 106) = 24.8$, $MS_e = 170$, $p < .0001$. This main effect was qualified by a Feedback × Item Difficulty interaction, $F(2, 105) = 4.0$, $MS_e = 137$, $p = .021$. Echoing earlier findings (Fischhoff, 1977), there was a significant hindsight effect on the hard and medium difficulty items, but not on the easy items. This may represent a ceiling effect on subjects' already high level of confidence in the correct answer. Regression analyses, where probabilities assigned to the correct alternative were regressed onto target norms (*t*), revealed a similar pattern and are shown in Figure 2. Feedback subjects had much higher intercepts, $F(1, 106) = 14.4$, $MS_e = 339$, $p < .0002$, while control subjects had marginally steeper slopes, $F(1, 106) = 3.6$, $MS_e = 0.08$, $p = .06$.

The hindsight effect was also present when considering the number of times that subjects in the control and feedback conditions selected the correct alternatives for each of the items. Feedback subjects more often indicated that they believed that they would have been able to select the correct alternative ($M = 72\%$) than was actually the case for control subjects ($M = 63\%$), $F(1, 106) = 17.1$, $MS_e = 0.04$, $p < .001$. As with the certainty ratings, the hindsight effect was more pronounced for more difficult items, $F(2, 105) = 6.0$, $MS_e = 0.03$, $p = .003$. Feedback subjects had particular difficulty recognizing that they would have selected the wrong answers to difficult items: feedback group, $M = 47\%$, and control group, $M = 30\%$. However, even though feedback subjects displayed hindsight on these difficult items, often "giving themselves the benefit of the doubt," it is still the case that they recognized they would have been incorrect in a majority (53%) of instances. The point is that feedback subjects were at least partially sensitive to the difficulties they would have experienced in foresight.

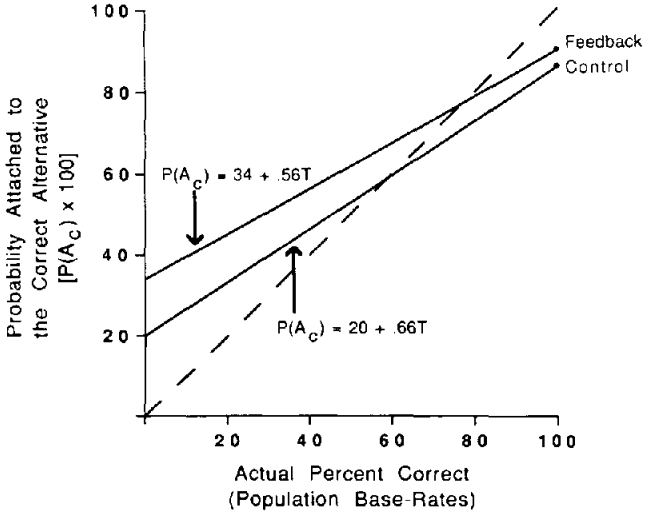

Figure 2. Hindsight effect in Experiment 1: Subjects' level of confidence in the correct alternative according to population base-rates.

Predictive accuracy analyses. The predictive accuracy analyses present quite a different picture of the effects of outcome feedback. Repeated measures MANOVAs of subjects' peer predictions (*p*) indicated a significant Feedback × Difficulty interaction, $F(2, 105) = 11.8$, $MS_e = 52.9$, $p < .001$. There were no differences in the peer predictions of feedback and control subjects on medium and easy items; however, on difficult items feedback subjects ($M = 51\%$) made predictions more in line with actual target norms (28%) than did controls ($M = 60\%$). This result was probed in more detail by calculating item by item absolute deviations ($|p - t|$). In general, feedback subjects ($M = 20.8$) were more accurate than control subjects ($M = 23.1$), $F(1, 106) = 5.1$, $MS_e = 48.1$, $p = .026$. But this main effect was qualified by a Feedback × Difficulty interaction, $F(2, 105) = 4.1$, $MS_e = 73.8$, $p = .02$, as feedback increased accuracy most on the difficult items.[3]

Individual level regression and correlation analyses were also conducted. Each subject's 20 peer predictions (*p*) were regressed onto the actual percent correct (*t*) for each item. On average, feedback subjects' peer predictions were more highly correlated [gamma(*t*, *p*)] with the actual target norms than were control subjects' predictions, $M = 0.37$ versus $M = 0.26$, $t(106) = 2.89$, $p < .005$. As is evident from the average regression lines for each group plotted in Figure 3, the feedback group's predictions are generally more accurate (closer to the dashed diagonal) than are those of the controls. The feedback group's intercept was significant lower, $t(106) = 3.15$, $p = .002$, while their slope coefficient was significantly greater, $t(106) = 3.14$, $p = .002$.[4]

[3] A similar analysis of absolute deviations was performed depending on whether subjects answered or, in the case of feedback subjects, said they would have answered a question correctly. Average absolute deviations were calculated separately for correct and incorrect items. Although feedback subjects made slightly larger absolute errors on correct responses ($M = 18.1$ vs. $M = 17.1$), they made much smaller absolute errors on incorrect problems ($M = 26.2$ vs. $M = 31.0$), a significant interaction, $F(1, 106) = 12.1$, $MS_e = 38.7$, $p < .001$, similar to the Feedback × Normative Item Difficulty reported previously.

[4] One objection voiced about the present results and those in the subsequent experiments was that the effects may be driven solely by the inclusion of questions that are misleading and therefore very difficult. It is true, especially in the two-alternative recognition task,

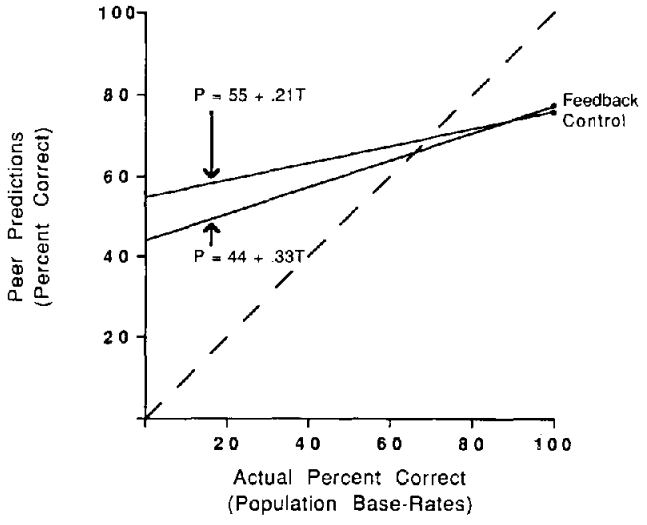

Figure 3. Information effect in Experiment 1: Average regression lines representing predictive accuracy (peer predictions regressed onto actual target norms).

Discussion

These results indicate that the provision of outcome feedback allowed subjects to tap into information that was useful in gauging normative item difficulty. This occurred despite the fact that outcome feedback contemporaneously produced a hindsight effect. It might seem strange that feedback increased accuracy most on the very items (difficult items more prone to commission errors) where subjects also displayed the most hindsight. The paradox can be explained as follows. Hindsight is extreme because subjects cannot appreciate exactly how overconfident they really would have been in prospect. Feedback is still informative, however, because it does moderate the extreme overconfidence that would have been present. Thus, despite the presence of hindsight bias, outcome feedback improved predictive accuracy in a related judgment domain, in this case peer (base-rate) predictions. It remains to be demonstrated that increases in accuracy resulted from an improved ability to discriminate between easy and difficult items.

Experiments 2 and 3

In Experiments 2 (E2) and 3 (E3), we employed a somewhat different paradigm in which alternative outcomes (answers) were not explicitly provided, but instead were generated by the subject. The paradigm in Experiment 1 was essentially a forced-choice recognition task. The procedure in E2 and E3 was an open-ended recall task, cued by the question, requiring subjects to generate answers based on retrieval from long-term memory. With this cued recall task we examine more closely the effect of outcome feedback on subjects' assessment of item difficulty, both for themselves and their peers.

A limitation of the recognition task employed in the first experiment is dependence on a half-range response scale. (This is also the case for a true–false procedure.) Such a scale constrains subjects' confidence in their own responses to lie between .5 and 1; any lower level of confidence would merit selection of the alternate response. However, predictions about the target population and postfeedback predictions about one's own response are not similarly constrained; it is quite possible that less than 50% of the target population could guess the correct answer or that, knowing the true answer, one recognizes that one had less than a 50% chance of guessing it. Hence, it is difficult to compare how well feedback and control subjects assessed their own level of knowledge, that is, calibration of own knowledge with population norms. This problem is eliminated in E2 and E3 by employing a recall paradigm in which all responses can lie between 0%–100%.

Unfortunately, the free-recall paradigm does not allow us to explicitly calculate a "pure" hindsight effect. In the cued recall task, control subjects were asked to generate an answer and to rate their confidence in the correctness of their answer. Feedback subjects were told the correct answer and then asked to indicate how likely it was that they would have been able to generate the correct answer. A pure measure of hindsight cannot be calculated in this task because control and feedback subjects are not necessarily rating the same piece of information. To measure pure hindsight, we would need control group ratings of the correct answer. This is easy to do in a recognition paradigm because if a subject guesses the wrong answer, $P(A_c)$ is simply 1 minus the probability placed on the wrong answer. However, in the recall task it is impossible to estimate $P(A_c)$ when control subjects answer questions incorrectly. Nevertheless, the data can still provide evidence concerning information effects net of hindsight. In E2 and E3 we assumed that there was some, albeit unspecified, hindsight effect. The key question was whether, despite hindsight, the information effect still emerges.

Method

Procedure. The experimental task utilized open-ended cued recall of general knowledge items. Control subjects saw 20 items of the form "What is the predominant religion of Albania?" First, control subjects generated a response to each question. They were instructed to write down an answer to each question even if they were "very unsure if [they were] right or wrong." Next they were asked to indicate how likely it was that their own answers were correct (o) by using an 11-point probability scale, where 0.0 was labeled *absolutely certain*

that misleading questions are the most difficult [i.e., low p(correct)], even more difficult than questions where the subject knows nothing (and therefore should guess correctly half the time). As we pointed out while laying out our model of outcome feedback, feedback *should* be more informative when commission errors are likely, because the subject learns not only what is right but also what is wrong. It is not the case, however, that the observed effects were found *only* because of misleading questions. In Experiment 1, we identified the five problems answered below chance [$P(C) < .4$ based on a 90% confidence level] and then redid all the analyses eliminating these items. A similar tactic, eliminating the five most difficult problems in each case, was used in the remaining experiments. Although the results tended to be less dramatic, no statistical inferences changed.

answer is incorrect, 0.50 was labeled *50–50 chance that answer is correct*, and 1.0 was labeled *absolutely certain answer is correct*. Finally, using a 0%–100% scale, subjects made predictions (p) about the percentage of their MBA peers who produced the correct answer (t).

Feedback subjects went through a similar procedure except that instead of generating answers to each question, they were shown the correct answers and told to study them carefully because they would be asked additional questions about these items. Next, for each item, feedback subjects indicated the probability that they would have produced the correct answer (o) assuming that they had not just seen the correct responses. Finally, they made the same peer predictions (p) as the controls.

Besides replicating the control and feedback conditions used in E2, in E3 we manipulated two other variables intended to influence the amount of hindsight subjects experience and then examined how this might affect the information value of feedback. The first manipulation, labeled *own-feedback-peer*, was intended to decrease hindsight and increase predictive accuracy. Subjects received feedback, but not until after answering each of the items and rating their own level of confidence; after the feedback, they then made peer predictions. We felt that hindsight might be reduced in this situation because subjects would have access to real rather than simulated retrieval attempts; and, if nothing else, "I-knew-it-all-along" reactions would be less likely to occur after a true retrieval failure. The second manipulation, labeled *feedback-delay-own-peer* was intended to increase the potential for hindsight and decrease predictive accuracy by introducing a delay between receipt of feedback and later judgment tasks. This condition was identical to the feedback condition except for the delay following outcome feedback. We reasoned that a delay between feedback and judgment might decrease the validity of later simulations of the retrieval experience. Fischhoff and Beyth (1975) found that longer delays between initial predictions and feedback increased the level of hindsight when subjects were asked to remember those predictions.

Stimuli. Two sets of 20 questions were prepared. The first set consisted of open-ended versions of the same questions as those used in Experiment 1. Although these stimuli do not represent a random sampling in any sense, we knew from Experiment 1 that they varied dramatically in terms of difficulty. Moreover, using the same questions allowed us to examine whether the type of task (recognition/recall) might make a difference. (In fact, we could discern no differences.) A second set of items was constructed to ensure that our results were not stimulus bound. The item set was selected from the 300 general knowledge questions developed by Nelson and Narens (1980) for use in feeling-of-knowing studies. They provided norms based on undergraduate subjects. To construct our set, we stratified the questions into deciles based on difficulty and then randomly chose two questions from each decile according to the undergraduate norms. Both stimulus sets were used in E2; only the first stimulus set was used in E3.

Subjects and design. Subjects were MBA students in the Graduate School of Business at the University of Chicago, $n = 158$ in E2 and $n = 81$ in E3. Both experiments were administered during regular classes. There was random assignment within each experiment. There were two independent variables in E2: outcome feedback (no feedback control or feedback) and stimulus set (two different versions). In E3 there were four conditions: control, feedback, own-feedback-peer, or feedback-delay-own-peer.

Results

For each stimulus set, the 20 open-ended questions were divided into three levels of difficulty: 7 difficult items (16% correct on average), 6 medium items (50% correct), and 7 easy items (73% correct).[5] The results obtained from the two stimulus sets were qualitatively indistinguishable, so the results from E2 are averaged across the two sets. Key results from the two experiments are summarized in Table 1.

Confidence in one's own knowledge. In E2, a repeated measures MANOVA of subjects' ratings of their own confidence (o) indicated that feedback subjects were less confident ($M = 37$) about difficult items than were controls ($M = 43$), whereas they were more confident ($M = 77$) about easy items than were controls ($M = 68$). This resulted in a significant Feedback × Item Difficulty interaction, $F(2, 153) = 17.5$, $MS_e = 193$, $p < .001$. The relevant comparison in E3 is between subjects making own confidence ratings without (or before) feedback (the control and own-feedback-peer groups) and those making own confidence ratings after feedback (the feedback and feedback-delay groups). The Feedback × Item Difficulty interaction was also significant in E3, $F(6, 150) = 3.8$, $MS_e = 190$, $p = .002$. In essence, the feedback seemed to increase subjects' confidence both in their ability to retrieve information that they probably did know (easy items) and in their inability to retrieve information that they most likely did not know (hard items).

This response polarization may have occurred because feedback provided subjects with information that was diagnostic of underlying base-rate item difficulty. For both studies, own confidence ratings made after receipt of feedback were better aligned with actual target norms [gamma(t, o) in Table 1] than were own confidence ratings without the benefit of feedback, $F(1, 154) = 35.9$, $MS_e = .03$, $p < .0001$, in E2, and $F(3, 77) = 6.6$, $MS_e = .03$, $p < .001$, in E3. The delay between feedback and own confidence ratings in E3 (delay-feedback-own-peer) decreased this correlation, though not significantly. At least for the 2-week delay used here, subjects were able to accurately remember or simulate likely prefeedback knowledge and couple it with outcome feedback to form a more accurate picture of item difficulty.

Figure 4 shows the average unstandardized regression analyses (regressing o onto t) for E2. The feedback group's intercept was lower than that of the controls, $F(1, 154) = 6.2$, $MS_e = 524$, $p = .014$; the feedback group also had steeper slopes on average, $F(1, 154) = 31.1$, $MS_e = 0.07$, $p < .0001$. Feedback had the same effects on the regression coefficients in E3, lower intercepts $F(3, 77) = 5.7$, $MS_e = 409$, $p = .001$, and steeper slopes $F(3, 77) = 8.3$, $MS_e = 0.09$, $p < .0001$. It appears that outcome feedback, besides resulting in the hindsight bias observed in Experiment 1, also leads to improved personal knowledge calibration in the sense that confidence levels are more closely aligned to population norms.

Predictive accuracy. The peer prediction results closely mimic those for own confidence ratings. As with previous research on knowledge projection (Nickerson et al., 1987;

[5] The target norms (t) were determined by the performance of control subjects in all the experiments. We also conducted all the analyses by using Nelson and Narens' (1980) undergraduate norms. The correlation between our MBA norms and the undergraduate norms was .83. The pattern of results held up in all respects except that the different measures of accuracy were slightly lower.

Table 1
Different Measures of Performance in Experiments 2 and 3

Condition	Agreement between confidence ratings and target norms			Level of projection	Accuracy of peer predictions (base-rate predictions)		
	gamma(t,o)	Intercept	Slope	gamma(p,o)	gamma(t,p)	Intercept	Slope
Experiment 2							
Control	.26	35	.46	.59	.22	44	.29
Feedback	.42	26	.70	.71	.38	29	.55
Experiment 3							
Control	.25[a]	36[b]	.41[a]	.63[a,b]	.34[a]	27	.49[a]
Feedback	.45[b]	19[a]	.79[b]	.73[b]	.43[b]	19	.65[b]
Own-feedback-peer	.26[a]	42[b]	.38[a]	.52[a]	.49[b]	26	.58[a,b]
Feedback-delay-own-peer	.37[b]	28[a,b]	.64[b]	.68[b]	.47[b]	19	.64[b]

Note. In Experiment 2, all differences between the control and feedback group are significant. In Experiment 3, within each variable, group means with different alphabetic superscripts are different from each other at the $p < .05$ level, using Newman-Keuls multiple comparison procedure.

Travers, 1943; Wood, 1978), subjects' estimates of what other people know are determined to a large extent by what subjects themselves know or think they know. Projection was operationalized as the correlation between subjects' own confidence ratings and their peer predictions [gamma(p, o)]. Knowledge projection was high for all groups, though comparable to other data concerning the relation between one's own position and predictions concerning one's peers (Hoch, 1987). The feedback groups, however, did display significantly higher levels of projection, relying on their own feelings of confidence more when making peer predictions, $F(1, 154) = 11.6$, $MS_e = 0.05$, $p < .001$, in E2, and $F(3, 77) = 5.5$, $MS_e = 0.03$, $p < .01$, in E3. It is possible that outcome feedback provides a systematic external source of information for both own and target judgments (see Hoch, 1988). Exposure to an external cue may reduce random error in subjects' responses and thereby increase the correlation between the two sets of judgments (p and o).

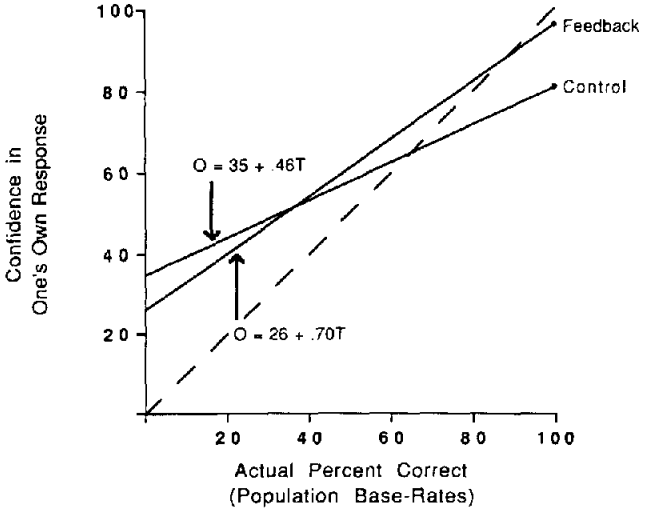

Figure 4. Average regression lines representing calibration of personal knowledge with population base-rates (own confidence ratings regressed onto actual target norms) in Experiment 2.

Repeated measures MANOVAS of peer predictions in E2 showed that feedback subjects' predictions were lower than those of controls on difficult items ($M = 37\%$ vs. $M = 48\%$) and slightly higher on easy items ($M = 69\%$ vs. $M = 66\%$), a significant Feedback × Item Difficulty interaction, $F(2, 153) = 18.3$, $MS_e = 96.8$, $p < .001$. This same interaction was not significant in E3. Analysis of item-by-item absolute deviations ($|p - t|$) indicated that in E2 the feedback subjects were more accurate than control subjects overall ($M = 23.5$ vs. $M = 27.3$), $F(1, 154) = 18.9$, $MS_e = 82.6$, $p < .0001$; however, this main effect was again qualified by a significant Feedback × Difficulty interaction, $F(2, 153) = 13.8$, $MS_e = 79.7$, $p < .001$. Most of the feedback group's improved accuracy was concentrated in the hard items. For E3, however, the absolute deviations showed only main effect of feedback and no interaction with item difficulty; absolute deviations of the control group ($M = 24.4$) were greater than those of the three feedback groups ($M = 21.7$), $t(77) = 2.05$, $p < .05$.

Correlation analyses also demonstrated greater predictive accuracy [gamma(t, p)] for the feedback groups, $F(1, 154) = 34.6$, $MS_e = 0.03$, $p < .0001$, in E2, and $F(3, 77) = 5.2$, $MS_e = 0.02$, $p < .003$, in E3. The unstandardized regression results for E2 are plotted in Figure 5. The regression line representing the feedback group is closer to the ideal (the underlying base-rates) at virtually every probability level, with significant differences in both intercept, $F(1, 154) = 28.0$, $MS_e = 301$, $p < .0001$, and in slope, $F(1, 154) = 44.5$, $MS_e = 0.06$, $p < .0001$. And although feedback had no effect on the intercepts in E3, it did increase the slopes, $F(3, 77) = 5.8$, $MS_e = 0.05$, $p < .001$. Receiving feedback *after* personally engaging in the recall task (own-feedback-peer) did not systematically increase accuracy above the regular feedback group.

Discussion

These two experiments provide additional evidence of the hypothesized multiple effects associated with outcome feedback. Even though subjects probably experienced feelings of hindsight after exposure to the correct answers, they also gained useful information from outcome feedback, enhancing the ability to discriminate between items on which they probably would have experienced a retrieval failure and those

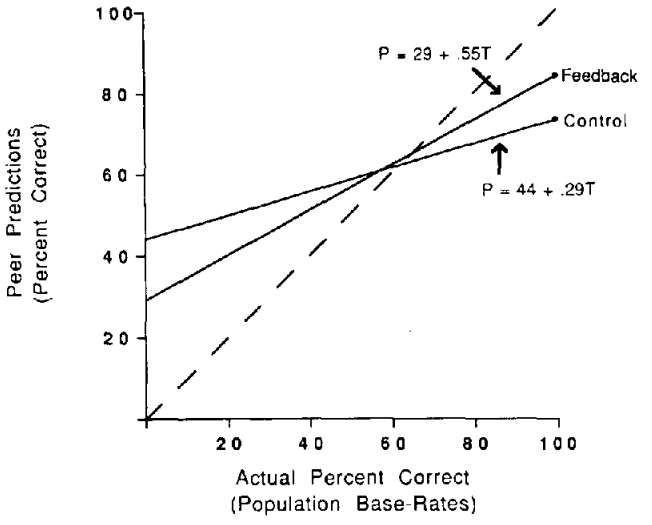

Figure 5. Information effect in Experiment 2: Average regression lines representing predictive accuracy (peer predictions regressed onto actual target norms).

on which they would have had a retrieval success. Thus, although feedback subjects might have exaggerated the probability they would have assigned to the correct answer in prospect, subjects were not completely overwhelmed by hindsight; instead, they were able to capitalize on the information contained in the feedback both to improve personal knowledge calibration and improve the accuracy of predictions concerning their peers. Whether information effects always emerge despite hindsight seems an open question, but Experiments 1–3 indicate that the information effect is quite robust.

We hypothesized that actually engaging in the recall task before receiving feedback might reduce hindsight, but we observed no increase in predictive accuracy. This suggests that the simulated retrieval efforts made by feedback subjects provide a reasonable approximation to the real retrieval attempts made by subjects unencumbered by outcome knowledge. Additional information (if any) gained from matching feedback to a real retrieval attempt did not translate into marginal improvements in accuracy.

We also hypothesized that temporal delay between receipt of outcome feedback and judgment might reduce accuracy by substantially increasing hindsight, but we found no such evidence. Possibly the delay was too short, not increasing hindsight enough to decrease accuracy. A future study might use a recognition task where hindsight could be calculated directly as a manipulation check. Alternatively, the item discrimination information that subjects abstract from the feedback might be encoded at the time of feedback, regardless of whether subjects are explicitly asked to use the information in making judgments. In processing the stimuli and the correct answers, subjects may incidentally encode familiarity or congruence with existing knowledge; later when asked to make judgments about the very same stimuli, they may be able to retrieve this information and use it to discriminate between items of varying difficulty.

Experiment 4

Our model of outcome feedback proposes that the ability of feedback subjects to experience differential surprise (diagnostic of underlying base-rates) on receipt of feedback drives the information effect. In this experiment both control and feedback subjects rated their level of surprise to each of the answers.

Method

Procedure. Both a recognition task (same as in Experiment 1) and a cued recall task (Experiments 2 and 3) were employed. The procedures for both the control and feedback groups were identical to those used earlier except for one additional task. After subjects completed the own confidence (o) and peer (p) rating tasks, all subjects (control and feedback) were shown the correct answers to each of the questions and asked to "indicate how surprised [they were] with the correct answer." Ratings were made using a 7-point scale, with 1 = *not at all surprised*, 4 = *moderately surprised*, and 7 = *extremely surprised*.

Subjects and design. Seventy-five MBA students participated during two regular class periods. The two independent variables were type of task (cued recall or forced-choice recognition) and outcome feedback (control or feedback). Subjects were randomly assigned to one of the four conditions.

Results and Discussion

The results replicated those in the previous studies. In the recognition task, the receipt of feedback led to hindsight. Feedback subjects assigned higher confidence ratings to the correct answer ($M = 71$) than did controls ($M = 63$), $F(1, 37) = 7.4$, $MS_e = 236$, $p = .01$. The Feedback × Item Difficulty interaction was not significant. Correlation analyses of the recall task showed that feedback subjects aligned their own confidence ratings more in line with target norms [gamma(t, o)] than did controls ($M = 0.43$ vs. $M = 0.19$), $F(1, 36) = 15.8$, $MS_e = 0.03$, $p < .001$. Correlation analyses of both the recall and recognition tasks indicated that feedback subjects made more accurate peer predictions [gamma(t, p)] than did the controls ($M = 0.37$ vs. $M = 0.17$), $F(1, 71) = 26.8$, $MS_e = 0.03$, $p < .0001$. Unstandardized regression analyses, where peer predictions (p) were regressed onto target norms (t), showed that feedback subjects had lower intercepts ($M = 33$ vs. $M = 49$), $F(1, 71) = 14.2$, $MS_e = 324$, $p < .001$, and steeper slopes ($M = 0.46$ vs. $M = 0.17$), $F(1, 71) = 37.1$, $MS_e = 0.04$, $p < .0001$.

Surprise analyses were conducted at the group and individual level. Five subjects (2 controls and 3 feedback) were excluded from the analyses because they indicated no surprise ($s = 1$) to all items. The average surprise ratings of the control ($M = 2.4$) and feedback ($M = 2.2$) groups did not differ, $F(1, 66) = 1.85$, $MS_e = 1.12$, $p = .18$. Average surprise ratings for each individual item were calculated separately for the four groups. The control and feedback groups provided similar surprise ratings; within tasks, the correlations between control and feedback ratings were greater than .8, ($p < .001$). This suggests that subjects experience similar surprise reactions regardless of whether they have access to an explicit attempt

to answer the question or a simulated attempt. This result is compatible with results in Experiment 3, where we found that the own-feedback-peer group was no more accurate than the regular feedback group, again suggesting that simulated retrieval attempts are diagnostic of explicit attempts.

We also examined the relation of surprise ratings to a crude measure of how surprised subjects *should* feel. For each question, we calculated the difference between the average confidence of the control group (o_c) and the objective target norms (t). When this difference is positive and large, it indicates an item where overconfidence is high and therefore a case where most subjects should be surprised at the correct answer. When this difference is near zero, it suggests that confidence levels are more in line with reality, and so most subjects would not be surprised. When the difference is negative and large (which happened only a few times), subjects on average are very underconfident and also might experience surprise at how obvious the correct answer turns out to be. To capture this relation, we calculated the correlation between average surprise ratings and the absolute difference between control group confidence ratings and target norms ($|o_c - t|$). For each of the four experimental groups, the correlations were greater than .8, ($p < .001$), again supporting the idea that the simulated retrieval attempts of the feedback group provide them with at least a partially valid sense of surprise.

The individual-level analyses, although statistically significant, were less convincing. For each subject, surprise ratings were correlated with the previously discussed ($|o_c - t|$) differences for each item. Although the feedback and control groups showed virtually identical results, the average correlations [gamma(s,$|o_c - t|$)] were low, an average of .24, ($p < .001$). We speculate that these lower individual-level correlations partly reflect response inconsistency and heterogeneity in use of the scale. There was a negative relation between surprise ratings and actual item difficulty, gamma(s, t) = $-.28$, $p < .001$, where difficult items were viewed as more surprising than easy items, with no differences among groups. But feedback subjects' confidence ratings and peer predictions were more highly correlated with surprise than were those of the control subjects. There was little connection between surprise ratings and p and o for the control group, gamma(s, p) = $-.09$ and gamma(s, o) = .15; because they did not know the answers, they could not experience surprise at the time they made their o and p ratings. For feedback subjects, in contrast, greater surprise led to lower o and p ratings, gamma(s, p) = $-.35$ and gamma(s, o) = $-.45$ (both $ps < .001$).

Experiment 5

This last study was conducted for two reasons. First, we were interested in the effects of information and hindsight in a domain other than general knowledge trivia tasks. Therefore, we provided outcome feedback about problems that required more extended deliberation. Second, we were still interested in whether it was possible to find situations where the balance between information and hindsight would be more likely to shift in the direction of hindsight. Therefore, we constructed tasks which, a priori, we believed would be more or less susceptible to hindsight bias and then examined the net effect of outcome feedback on predictive accuracy.

One type of problem that seems especially prone to hindsight is the heterogeneous class of tasks known as insight problems (Bourne, Ekstrand, & Dominowski, 1971; Maier, 1970; Weisberg & Alba, 1981). Although there is no exact agreement as to what is and what is not an insight problem, there is a general consensus that insight problems often are solved very suddenly, a "flash" experience accompanied by an "aha" response. The Gestalt principle of functional fixedness characterizes insight problems as those where subjects potentially have access to the correct answer but are momentarily "blocked" from recognizing it by other more obvious incorrect solutions. The solutions to insight problems are usually not complicated or exotic; they often seem fairly obvious once they are solved or the answer is revealed (which may be why authors typically make their readers turn to another page to find the answer). Once the answer is known (through feedback), subjects may find it difficult to simulate all the incorrect solutions that might have been attempted.

Metcalfe (1986a) found that although subjects were quite accurate in their feeling-of-knowing judgments concerning long-term memory phenomena, accurate metacognitions about insight problems were almost nonexistent. Her subjects demonstrated little insight about their ability to solve insight problems. In another series of studies, Metcalfe (1986b) found that subjects had great difficulty anticipating the onset of insight. Her article ends with this statement:

> To conclude with the practical question of whether one should believe people when they say that they have almost got the answer to a difficult insight problem, results of the present series of experiments suggest that we should be wary: At least with the class of problems studied here, premonitions of insight predict mistakes. (p. 634)

The cues that subjects use to simulate memorability when recall is prevented (either by recall failure in feeling-of-knowing studies or by preemptive outcome feedback in our experiments) seem to be absent for insight problems. Thus, feedback about insight problems may not provide subjects with much diagnostic information.

To provide a contrast for the insight problems, we developed problems that were more incremental in nature and not solved suddenly. Unlike insight problems, where the answer often seems to "spring" into the subjects' consciousness very suddenly, our noninsight problems required subjects to work toward a goal in smaller steps. In principle, these incremental problems were designed so that subjects could monitor the amount of progress they had made.

Method

Stimuli and procedure. Each subject tried to solve four problems: two insight and two incremental problems. The insight problems were the "chain problem" (Wickelgren, 1974), which required subjects to connect four small chains into one large chain in a prespecified number of moves; and the "hats problem" (Adams, 1974), a variant of many common deductive reasoning tasks. Two incremental tasks, both requiring generation of multiple answers, were designed for this experiment. One problem was anagram-like, requiring subjects to generate 20 different words out of the letters DECISION MAKING; the other problem required subjects to name 12 different African countries.

Subjects were given 3 min to solve each problem. After working on each problem, they made two judgments. First, they predicted the percent of their college peers who would solve the task in the allotted 3 min. Second, they rated how difficult they found the problem on a 10-point scale, where 1 = *very difficult* and 7 = *very easy*. The order of the problems was counterbalanced. In the feedback condition, subjects were provided with outcome feedback after attempting to solve each problem but before making the two judgments. Because there were no unique solutions to the incremental problems, the feedback consisted of one example of the many answers that were possible (e.g., one list of 12 countries out of a possible 50+).

Subjects and design. Sixty-four undergraduates at the University of Chicago were each paid $5 to participate. The task took about 30 min to complete and was embedded within other unrelated tasks. Outcome feedback was manipulated between subjects, while type of problem (insight or incremental) and replication (first and second problems) were manipulated within subjects. The counterbalancing order of presentation variable had no effect, so the data were collapsed across this variable.

Results

There were no differences in solution rates for the feedback (29%) and control (31%) groups. This was expected because both groups were functionally equivalent during the problem-solving stage of the experiment. Target peer predictions were analyzed by using a 2 × 2 × 2 Feedback × Problem Type × Replication repeated measures MANOVA including both solvers and nonsolvers.[6] Figure 6 shows the significant Feedback × Problem Type interaction, $F(1, 62) = 11.68$, $MS_e = 478$, $p < .001$. On insight problems, exposure to the correct answer caused subjects to estimate that a larger fraction of their peers would successfully solve the problem. The reverse pattern is apparent for incremental, noninsight problems. For noninsight problems, outcome feedback led to a downward revision in estimates of the fraction of peers who would provide a correct solution. On average, subjects in all conditions overestimated the performance of their peers. However, although outcome feedback improved predictive accuracy for incremental problems, finding out the right answer to insight problems was actually detrimental in terms of predictive accuracy. The data concerning subjects' subjective judgments of problem difficulty provided a similar though weaker pattern, $F(1, 62) = 2.52$, $MS_e = 4.24$, $p = .12$.

Discussion

Experiments 1–4 showed that outcome feedback can often provide useful diagnostic information, but the results of Experiment 5 indicate that there are situations where feedback may not be as helpful and may actually be detrimental. Although the results are far from definitive, they do suggest that outcome feedback about certain stimuli can elicit extreme hindsight that overrides any potential information. In the case of insight problems, subjects had difficulty appreciating how easy it would have been to pursue dead-end solutions and become fixated on unwarranted assumptions. Subjects may simply not experience an appropriate level of surprise when they see the right answers to difficult insight problems because, unlike long-term memory phenomena where people display valid feeling-of-knowing, subjects appear to have difficulty simulating the insight problem-solving experience. "Knew it all along" feelings may be so strong that they eliminate the possibility of using surprise reactions to discriminate between easy and difficult items. Feedback may seduce subjects into an unjustified sense of problem closure, the result being strong hindsight that causes bias and undermines the information value of feedback.

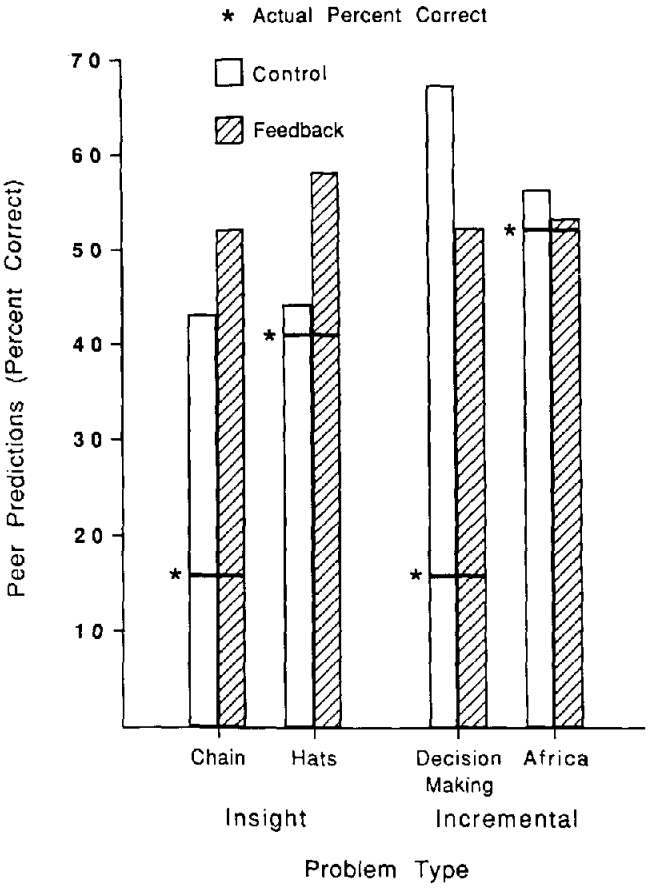

Figure 6. Predictions of peer performance on insight and incremental problems in Experiment 5.

General Discussion

This article identifies clear limits to the influence of hindsight on other aspects of judgment. Outcome feedback can have multiple effects, both functional and disfunctional. In many situations, subjects extract diagnostic information from outcome feedback, despite the fact that such feedback leads them to overestimate how much they would have known in foresight. Our research shows that hindsight does not preclude surprise, even though it does lead subjects to be less surprised than they should be. The surprise reactions that subjects do experience, albeit insufficient, are nevertheless often diagnos-

[6] The data were also analyzed by excluding all subjects who solved the problems. Solvers typically provided much higher estimates of solution rates as they projected from their own successful experience. Using univariate analyses of each problem, similar differences between control and feedback subjects emerged when only nonsolvers were considered.

tic. Our data indicate that subjects are capable of using these internal surprise reactions as a basis for discriminating underlying base-rate difficulty and probability of occurrence even in the presence of hindsight. In Experiments 1–4, feedback increased the accuracy of peer predictions by over 150% (on an R^2 basis).

Throughout psychology, accurate outcome feedback is recognized as the most important determinant of adaptivity and rate of learning (Anderson, 1983; Brunswik, 1952; Einhorn & Hogarth, 1978; Hogarth, 1981; Nelson, 1971). Without feedback, how can we possibly learn effective judgment and decision making policies? What our research and previous research on hindsight demonstrate is that outcome feedback is a necessary though not sufficient condition for adaptive learning to take place. When hindsight is so strong as to obliterate any potential for surprise reactions, then feedback may actually inhibit adaptation. However, it appears that there are a variety of situations where subjects suffer only moderate hindsight while concurrently extracting diagnostic information from outcome feedback.

The hindsight effect is applicable to situations where people attempt to gauge the likelihood of specific outcomes. Our findings are consistent with earlier research suggesting that outcome information leads to exaggerated retrospective estimates of the likelihood that oneself or other people would predict what one knows has occurred. The information effect, in contrast, applies to judgments of outcome base-rate—for example, judgments of whether oneself or others would correctly answer a question or anticipate an eventual outcome. There are many situations in which such information is valuable. For example, consider a trial; we may be concerned with whether the jurors correctly answer the question of guilt or innocence, whatever it may be. How will inside information about the defendant's guilt affect our ability to predict the accuracy of the jury? When selecting a doctor, we are concerned not with the likelihood that he will diagnose "cancer" but with whether the diagnosis is correct. Or suppose you are deciding whether to invest in a company. The company's success depends on whether it invests in the right technology from among a number of alternatives. Who would better be able to predict whether the company will make the right choice: a market analyst who himself does not know which technology is right or a scientist who does? Finally, who can better predict whether Professor X will find his way to the party: the host, who knows the correct directions, or a guest who has only a vague idea?

Several lines of future research are suggested by our work. First, we need to develop a better understanding of the kinds of tasks that are especially prone to strong hindsight, no-surprise reactions. The strong hindsight effects found on insight problems suggest that subjects' lack of awareness of the potential for "garden path" errors (Johnson, Moen, & Thompson, in press) and unwarranted assumptions due to backward reasoning are two such factors that deserve further study. More generally, we need to gain a more comprehensive understanding of the limits of the information extracted from feedback. Although we found that subjects could improve predictive accuracy on related tasks after feedback, it is not clear how far these information effects might carry. Lichtenstein and Fischhoff (1980) found that extensive training involving summary feedback dramatically improved subjective probability calibration on related tasks; however, they also found that transfer to other task domains was limited. Our data provide no indication of the extent to which information effects generalize or how long they endure. For example, although overconfidence is moderated on the items for which subjects receive feedback even in the presence of hindsight, it is not clear how hindsight might influence confidence at a more global level. The existence of hindsight indicates that subjects have not adequately acknowledged how much feedback has taught them, and, over time, continued underestimation of what has been learned from feedback may unjustifiably increase metacognitions about one's level of competence.

In the final analysis, it seems that we have to take the good (information) with the bad (hindsight) when it comes to outcome feedback. The picture is not as bleak as previous research on hindsight has suggested because that which leads to hindsight also leads to information. A more complete understanding of the interplay between hindsight and information, however, remains the province of future research. When this research is done, it will be interesting to see whether the reaction is "I knew it all along" or "I never would have guessed it" or a mixture of the two.

References

Adams, J. L. (1974). *Conceptual blockbusting.* Reading, MA: Addison-Wesley.

Anderson, J. R. (1983). *The architecture of cognition.* Cambridge, MA: Harvard University Press.

Arkes, H. R., Wortmann, R. L., Saville, P. D., & Harkness, A. R. (1981). Hindsight bias among physicians weighting the likelihood of diagnoses. *Journal of Applied Psychology, 66,* 252–254.

Atkinson, R. C., & Juola, J. F. (1974). Search and decision processes in recognition memory. In D. H. Krantz, R. C. Atkinson, R. D. Luce, & P. Suppes (Eds.), *Contemporary developments in mathematical psychology: Vol. 1. Learning, memory, & thinking.* San Francisco: Freeman.

Blake, M. (1973). Prediction of recognition when recall fails: Exploring the feeling-of-knowing phenomenon. *Journal of Verbal Learning and Verbal Behavior, 12,* 311–319.

Bourne, L. E., Jr., Ekstrand, B. R., & Dominowski, R. L. (1971) *The psychology of thinking.* Englewood Cliffs, NJ: Prentice-Hall.

Brown, R., & McNeill, D. (1966). The "tip of the tongue" phenomenon. *Journal of Verbal Learning and Verbal Behavior, 5,* 325–337.

Brunswik, E. (1952). The conceptual framework of psychology. In *International Encyclopedia of Unified Science* (Vol. 1, No. 10). Chicago: University of Chicago Press.

Camerer, C., Loewenstein, G., & Weber, M. (in press). The curse of knowledge in market settings: An experimental analysis. *Journal of Political Economy.*

Campbell, J. D., & Tesser, A. (1983). Motivational interpretations of hindsight bias: An in-individual difference analysis. *Journal of Personality, 51,* 605–620.

Dellarosa, D., & Bourne, L. E. (1984). Decision and memory: Differential retrievability of consistent and contradictory evidence. *Journal of Verbal Learning and Verbal Behavior, 23,* 669–682.

Einhorn, H. J., & Hogarth, R. M. (1978). Confidence in judgment: Persistence of the illusion of validity. *Psychological Review, 85,* 395–416.

Einhorn, H. J., & Hogarth, R. M. (1981). Behavioral decision theory:

Processes of judgment and choice. *Annual Review of Psychology, 32*, 53–88.

Fischhoff, B. (1975). Hindsight ≠ foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance, 1*, 288–299.

Fischhoff, B. (1977). Perceived informativeness of facts. *Journal of Experimental Psychology: Human Perception and Performance, 3*, 349–358.

Fischhoff, B. (1982). For those condemned to study the past: Heuristics and biases in hindsight. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 335–351). New York: Cambridge University Press.

Fischhoff, B., & Beyth, R. (1975). "I knew it would happen"—Remembered probabilities of once-future things. *Organizational Behavior and Human Performance, 13*, 1–16.

Gardiner, J. M., & Klee, H. (1976). Memory for remembered events: An assessment of output monitoring in free recall. *Journal of Verbal Learning and Verbal Behavior, 15*, 227–233.

Gentner, D., & Collins, A. (1981). Studies of inference from lack of knowledge. *Memory & Cognition, 9*, 434–443.

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review, 91*, 1–67.

Glucksburg, S., & McCloskey, M. (1981). Decisions about ignorance: Knowing that you don't know. *Journal of Experimental Psychology: Human Memory and Learning, 7*, 311–325.

Hart, J. T. (1965). Memory and the feeling-of-knowing experience. *Journal of Educational Psychology, 56*, 208–216.

Hasher, L., Attig, M. S., & Alba, J. W. (1981). I knew it all along: Or, did I? *Journal of Verbal Learning and Verbal Behavior, 20*, 86–96.

Hoch, S. J. (1985). Counterfactual reasoning and accuracy in predicting personal events. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*, 719–731.

Hoch, S. J. (1987). Perceived consensus and predictive accuracy: The pros and cons of projection. *Journal of Personality and Social Psychology, 53*, 221–234.

Hoch, S. J. (1988). Who do we know: Predicting the interests and opinions of the American consumer. *Journal of Consumer Research, 5*, 315–324.

Hogarth, R. M. (1981). Beyond discrete biases: Functional and dysfunctional aspects of judgmental heuristics. *Psychological Bulletin, 90*, 197–217.

Johnson, P. E., Moen, J. B., & Thompson, W. B. (in press). Garden path errors in diagnostic reasoning. In L. Bole & M. J. Coombs (Eds.), *Computer expert systems*. Amsterdam: Springer-Verlag.

Juola, J. F., Fischler, I., Wood, C. T., & Atkinson, R. C. (1971). Recognition time for information stored in long-term memory. *Perception & Psychophysics, 10*, 8–14.

Klee, H., & Gardiner, J. M. (1976). Memory for remembered events: Contrasting recall and recognition. *Journal of Verbal Learning and Verbal Behavior, 15*, 471–478.

Koriat, A., & Leiblich, I. (1974). What does a person in a "TOT" state know that a person in a "don't know" state doesn't know? *Memory & Cognition, 2*, 647–655.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Memory and Learning, 6*, 107–118.

Krinsky, R., & Nelson, T. O. (1985). The feeling of knowing for different kinds of retrieval failures. *Acta Psychologica, 58*, 141–158.

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? The calibration of probability judgments. *Organizational Behavior and Human Performance, 20*, 159–183.

Lichtenstein, S., & Fischhoff, B. (1980). Training for calibration. *Organizational Behavior and Human Performance, 26*, 149–171.

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). New York: Cambridge University Press.

Maier, N. R. F. (1970). *Problem solving and creativity in individuals and groups*. Belmont, CA: Brooks/Cole.

Mandler, G. (1972). Organization and recognition. In E. Tulving & W. Donaldson (Eds.), *Organization of memory* (pp. 139–166). New York: Academic Press.

Mandler, G. (1980). Recognizing: The judgment of previous occurrence. *Psychological Review, 87*, 252–271.

Metcalfe, J. (1986a). Feeling-of-knowing in memory and problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12*, 288–294.

Metcalfe, J. (1986b). Premonitions of insight predict impending error. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12*, 623–534.

Mitchell, T. R., & Kalb, L. S. (1981). Effects of outcome knowledge and outcome valence on supervisors' evaluations. *Journal of Applied Psychology, 66*, 604–612.

Nelson, T. O. (1971). Extinction, delay, and partial-reinforcement effects in paired-associate learning. *Cognitive Psychology, 2*, 212–228.

Nelson, T. O., Gerler, D., & Narens, L. (1984). Accuracy of feeling-of-knowing judgments for predicting perceptual identification and relearning. *Journal of Experimental Psychology: General, 113*, 282–300.

Nelson, T. O., R. J. Leonesio, R. S. Landwehr, & L. Narens (1986). A comparison of three predictors of an individual's memory performance: The individual's feeling-of-knowing versus normative feeling-of-knowing versus base-rate item difficulty. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12*, 279–288.

Nelson, T. O., & Narens, L. (1980). Norms for 300 general-information questions: Accuracy of recall, latency of recall, and feeling-of-knowing ratings. *Journal of Verbal Learning and Verbal Behavior, 19*, 338–368.

Nickerson, R. S., Baddeley, A., & Freeman, B. (1987). Are people's estimates of what other people know influenced by what they themselves know? *Acta Psychologica, 64*, 245–259.

Reder, L. M. (1982). Plausibility judgments versus fact retrieval: Alternative strategies for sentence verification. *Psychological Review, 89*, 250–280.

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

Travers, R. M. W. (1943). A study of the ability to judge group-knowledge. *American Journal of Psychology, 56*, 54–65.

Weisberg, R. W., & Alba, J. W. (1981). An examination of the alleged role of "fixation" in the solution of several "insight" problems. *Journal of Experimental Psychology: General, 110*, 169–192.

Wickelgren, W. A. (1974). *How to solve problems*. San Francisco: Freeman.

Wood, G. (1978). The knew-it-all-along effect. *Journal of Verbal Learning and Verbal Behavior, 4*, 345–353.

Yaniv, I., & Meyer, D. E. (1987). Activation and metacognition of inaccessible stored information: Potential bases for incubation effects in problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*, 187–205.

Received January 18, 1988
Revision received September 26, 1988
Accepted October 4, 1988 ■