Organizational Behavior and Human Decision Processes
Vol. 83, No. 2, November, pp. 282–309, 2000
doi:10.1006/obhd.2000.2910, available online at http://www.idealibrary.com on IDEAL®

Training to Improve Calibration and Discrimination: The Effects of Performance and Environmental Feedback

Eric R. Stone and Ryan B. Opel

Wake Forest University

This study investigated whether calibration and discrimination are distinct or related aspects of probability judgment accuracy by examining the effects of two different training techniques. Participants received either performance feedback or environmental feedback, and we measured their improvement in calibration and discrimination as a function of feedback type. Whereas performance feedback reduced participants' overconfidence and environmental feedback improved discrimination, neither type of feedback led to an improvement on the other component. In fact, environmental feedback led to an increase in overconfidence. We take these results as evidence that calibration and discrimination are dissociable abilities that require separate training techniques for improvement.

Researchers in the field of judgment and decision making have long tackled the question of how good people are at judgment tasks. As early as 1923, Wallace demonstrated that corn judges' predictions of relative yield correlated only .2 with the actual yield. Since then, perhaps surprisingly, a number of related findings have emerged suggesting deficiencies in the judgment process of professionals in such disparate fields as medicine (cf. Ayton, 1992; Bolger &

This experiment was presented as part of a poster session at the annual meeting of the Society for Judgment and Decision Making, Dallas, November 1998. We thank Bernadine Barnes for her helpful instruction in art history, Martine L. Sherrill and Christine K. Flory for their assistance and patience in assembling the numerous art slides, the Wake Forest University Art Department for loaning us the art slides, and Chantal M. Poister for her assistance in conducting the research. Also, we thank Richard R. Hoffman, Janine M. Jennings, Mark V. Pezzo, William Fleeson, and Catherine E. Seta for helpful discussions on the ideas incorporated in this paper. Ryan Opel is now at the School of Law and Department of Psychology, Duke University.

Address correspondence and reprint requests to Eric R. Stone, Department of Psychology, Wake Forest University, Box 7778 Reynolda Station, Winston-Salem, NC 27109. E-mail: estone@wfu.edu.

0749-5978/00 $35.00

Wright, 1992; Eddy, 1982; Lichtenstein, Fischhoff, & Phillips, 1982; Wallsten & Budescu, 1983; Wigton, 1988; Yates, 1990), business (cf. Bolger & Wright, 1992; Chan, 1982; Kahneman & Lovallo, 1993; Wallsten & Budescu, 1983; Yates, 1990), and clinical psychology (cf. Chan, 1982; Garb, 1989; Smith & Dumont, 1997). Similar findings have been obtained with laypeople making judgments in a variety of different contexts (cf. Lichtenstein et al., 1982; Yates, 1990). There have, however, been notable exceptions to this pessimistic conclusion, with results varying both as a result of the domain under investigation and as a result of the measure of judgmental accuracy used (see, e.g., Ayton, 1992; Yates, 1990).

Given the generally pessimistic but inconsistent results regarding the judgmental ability of experts, a number of researchers have attempted to determine whether or not people can be trained to make better judgments. If people's judgment processes can be improved via training, this would have obvious applied benefits, as these types of procedures could be incorporated into more formalized education. Further, by focusing on different measures of judgmental ability, training studies also provide insight into determining which aspects of the judgment process are related and which ones are distinct. Similar to the logic behind research on dissociation in memory (see, e.g., Payne & Wenger, 1998), research showing that a training technique improves one aspect of judgmental accuracy but not another provides evidence that those two aspects of judgment are distinct. This knowledge, in turn, could be used in the development of judgment aids directed at those aspects of the judgment process that are particularly weak and to training techniques designed specifically to produce expertise in that particular aspect of judgmental accuracy. Ultimately, it should be possible to determine what has led to the good performance of experts that has been found in some previous research (e.g., Keren, 1987; Murphy & Winkler, 1984) and to employ these procedures in a broad range of domains.

The focus of the present paper is on the two most studied aspects of probability judgment, calibration and discrimination. In particular, we examine whether these aspects require related or distinct skills by determining whether training procedures that affect one of these skills have any effect on the second measure. We shall proceed by discussing these different aspects of probability judgment, then discuss two types of feedback that serve as common training techniques, and conclude by discussing our present research.

ASPECTS OF JUDGMENTAL ACCURACY FOR PROBABILITY JUDGMENTS OF DISCRETE EVENTS

Much of the work on judgmental accuracy has focused on judgments of likelihood, in order to allow the judge to express uncertainty in his or her opinion. Primarily for reasons of simplicity, these types of tasks have tended to focus on situations where the person is judging between two discrete alternatives. Specifically, respondents are asked if a certain event will occur or not and to provide a probability judgment as to the likelihood of occurrence of that event. For example, in a study by Christensen-Szalanski and Bushyhead (1981),

physicians judged the probability of pneumonia on the basis of a physical examination. The event to which the probability judgment is assigned is referred to as the target event, which in this case was the occurrence of pneumonia. Perhaps the most common evaluation criterion is the probability score (PS), and is calculated via $(f - d)^2$, where f is the judgment of the likelihood of the target event's occurrence (e.g., the stated probability of pneumonia), and d is the actual outcome (coded as 0 if the target event does not occur and 1 if the target event does occur). A probability score is calculated for each of many judgments (say, 50 different estimates of the likelihood of pneumonia), and these are averaged to form the mean probability score ($\overline{PS}$), also known as the Brier score (Brier, 1950). The lower the $\overline{PS}$, the better one's judgments.

The mean probability score provides an estimate of overall judgmental accuracy, but in itself provides no information for *why* a judge might perform well or poorly; indeed, it is very possible for two sets of judges to achieve, on average, the same $\overline{PS}$, but for one set of judges to be superior in some aspects of the judgment process but inferior in others (see, e.g., Yates et al., 1989). Judgment researchers have long recognized this fact and have thus decomposed the $\overline{PS}$ into more useful accuracy measures. It can be shown (see, e.g., Murphy, 1973; Yates, 1982, 1994) that the $\overline{PS}$ is a composite of three components: calibration, discrimination, and the variability in the target event.[1] In the typical judgment situation, the judge is asked to report probability estimates, usually in terms of judgment categories (e.g., a 0 chance of pneumonia, a .1 chance, a .2 chance, etc.). Calibration refers to the match between these judgment categories and the percentage of times that the target event actually occurs. There are many ways in which a judge could exhibit poor calibration, but the most commonly discussed one is overconfidence, where the judge overestimates his or her likelihood of being correct. Discrimination (also known as resolution) refers to a judge's ability to discriminate between situations when the target event occurs and when it does not occur. Good discrimination is thus achieved when the judge's forecast categories produce conditional base rates that are far from the overall base rate of the event's occurrence (Braun & Yaniv, 1992). The third component, variability, is not controllable by the judge and thus is not an accuracy measure in the sense the other two components are, as long as the target event is externally defined. We discuss this issue in more detail later in the introduction.

It is important to recognize that the concepts of calibration and discrimination are distinct, in the sense that a judge could do well on one aspect of judgmental accuracy but poorly on the other. As one example, Liberman and Tversky (1993) discuss a hypothetical case of a doctor estimating the sex of a newborn infant with probability .50. If this doctor were to make this judgment repeatedly, he or she would be perfectly calibrated, as infants are male half the time and female the other half. However, the doctor's judgments do not discriminate between situations where the infant is male from when the infant

[1] See Yates (1994) for the actual decomposition and the formulas associated with the calibration and discrimination concepts.

is female. Conversely, a doctor could make judgments that do help to discriminate between two possible outcomes, even if he or she is highly over (or under) confident.

To see this, it is often helpful to examine a calibration graph, which depicts people's performance graphically (see Fig. 1). The abscissa includes the judge's probability judgment categories and the ordinate indicates the percentage of times the target event occurred for each judgment category. The diagonal line represents perfect calibration, as points on that line indicate that the judge is perfectly calibrated for each of his or her probability judgment categories (judgments of .20 occur 20% of the time, etc.). Points far from the base rate (in this case, 50%) on the ordinate indicate good discrimination, as they discriminate between situations where the target event occurs from when it does not occur. Figure 1a depicts the previously-discussed situation from Liberman and Tversky, where the doctor always provided .50 probability judgments that the newborn will be male. It should be clear that the judge is perfectly calibrated (his or her judgments fall on the line of perfect calibration), yet has nil discrimination. Conversely, Fig. 1b illustrates a situation where the doctor is poorly calibrated (specifically, overconfident), yet shows better discrimination. The overconfidence is indicated by the fact that a line connecting the points on the calibration graph would be too horizontal, in particular, that when the doctor made extreme judgments (e.g., there is a 100% chance the newborn will be male) he or she was often incorrect. However, this doctor does have better discrimination, as the judgment categories do help determine what the sex of the child will be (e.g., there is a much higher probability that the newborn will be male when the doctor makes a judgment of 1.0 than when a judgment of 0.0 is made).

TRAINING PROCEDURES TO IMPROVE CALIBRATION AND DISCRIMINATION

As discussed previously, training studies that focus on distinct aspects of the judgment process are particularly important, as they indicate what aspect of the judgment process is lacking. Along the same lines, by examining the relative impact of specific training techniques on different accuracy components (e.g., calibration and discrimination), it is possible to determine whether those components are dissociable. In other words, are the abilities to discriminate between possible outcomes and to provide well-calibrated probability judgments separate or related abilities?

Note that other authors have approached similar problems by constructing theoretical models of the judgment process and relating the issues under investigation to the particular theoretical model (see, e.g., McClelland & Bolger, 1994; Wallsten, Bender, & Li, 1999; Wallsten & González-Vallejo, 1994). Clearly, advancing our understanding of how judgments are made is a worthwhile goal. The focus of the present paper, however, is on the skills necessary for good calibration and discrimination, regardless of how the judgments are constructed. Theoretically, this translates into determining whether these abilities

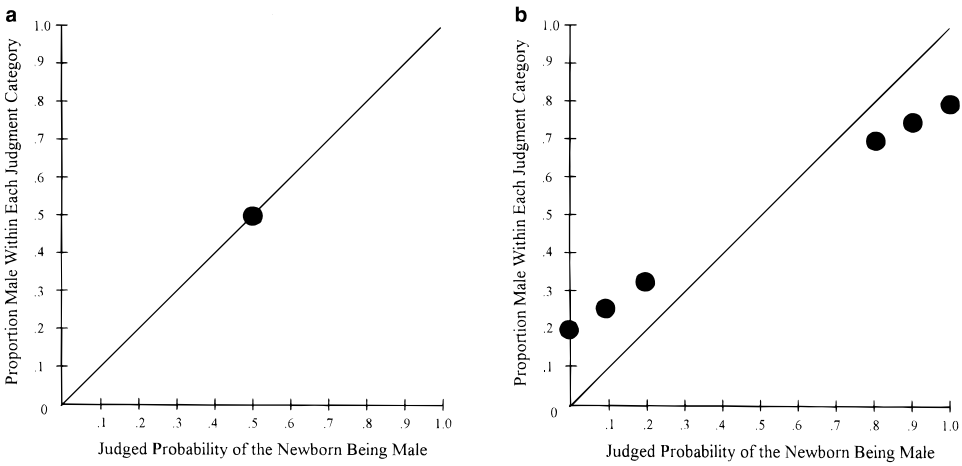

FIG. 1. Two hypothetical calibration graphs for the judgment of the sex of a newborn: (a) Doctor A: Good calibration, poor discrimination; (b) Doctor B: Good discrimination, poor calibration.

are dissociable. From a more applied perspective, the issue is whether educational training techniques that target one accuracy component (calibration or discrimination) will have any impact on the other component or whether separate programs need to be developed to target both components.

To answer these questions, it is necessary to distinguish among different types of training techniques. To this end, Benson and Önkal (1992) discuss a particularly useful distinction between different types of feedback (a type of training) that a person could receive. By examining these types of feedback and their effects on calibration and discrimination, it should be possible to determine if these training procedures have different or similar effects on the two aspects of judgmental accuracy. For the present purposes, we are focusing on two of these types of feedback: performance feedback and environmental feedback. Performance feedback involves providing information about the accuracy of one's judgments in general, for example, that the person was overconfident or that he or she achieved good discrimination. Environmental feedback involves providing information about the event to be predicted. Although this term has been used somewhat differently by different researchers, we are using it to refer to any domain-specific information given about the task under consideration. For example, informing a doctor that a high fever is associated with pneumonia would be a form of environmental feedback, because it provides information about the task (predicting pneumonia) rather than about the judgments made by the doctor. Note that both performance and environmental feedback are distinct from outcome feedback, i.e., information as to whether or not a particular judgment was correct.

We shall proceed by reviewing the work that has been done with performance feedback and environmental feedback, arguing that there is reason to believe that performance feedback improves calibration but not discrimination and that environmental feedback improves discrimination but not calibration. We

will then discuss a study that confirms these hypotheses and end by relating the results to a more general theory of expertise.

PERFORMANCE FEEDBACK

A number of researchers have suggested that performance feedback is a particularly effective method for improving calibration (see, e.g., Fischhoff, 1982). Perhaps the most intensive study using performance feedback was conducted by Lichtenstein and Fischhoff (1980). Subjects completed 11 training sessions of 200 two-alternative general knowledge questions. At the completion of each training session, they were given personalized performance feedback, including calibration graphs as well as specific performance measures such as $\overline{PS}$, calibration, and overconfidence. This feedback was then discussed with the subject for 5 to 10 minutes. There was a clear improvement in calibration as a result of the training. Most impressively, it appears as if the success of the feedback resulted primarily from the initial training session. These results led Lichtenstein and Fischhoff to conclude that one intensive performance feedback session is sufficient for a person to become well calibrated. Similar results have been found for a range of different tasks, at least those involving judgments of discrete events (e.g., Adams & Adams, 1958; Benson & Önkal, 1992; Bornstein & Zickafoose, 1999; Oskamp, 1962; Sharp, Cutler, & Penrod, 1988).

Thus, a great deal of research has suggested that performance feedback improves people's calibration abilities. There is good reason, however, to expect that performance feedback without accompanying outcome feedback will not improve discrimination, as performance feedback provides no information to help determine whether or not a particular event will occur (Benson & Önkal, 1992; see also Yates, 1994). We are aware, however, of only a limited number of studies that have examined the effect of performance feedback on discrimination per se (Benson & Önkal, 1992; Lichtenstein & Fischhoff, 1980; Sharp et al., 1988). In one such study, Benson and Önkal (1992) asked participants to judge the outcome of football games and provided some of their participants with performance feedback. Additionally, they provided one group of participants feedback on their calibration and another group of participants feedback on their discrimination scores (both types of performance feedback). As in the previously-discussed Lichtenstein and Fischhoff (1980) study, performance feedback regarding calibration improved participants' calibration scores, primarily by decreasing the level of overconfidence. More importantly, however, there was no impact of providing calibration feedback on discrimination in either the Benson and Önkal (1992) or the Lichtenstein and Fischhoff (1980) study. Similarly, those participants given performance feedback regarding discrimination in Benson and Önkal's study did not improve on any measure. Thus, Benson and Önkal concluded that "our results suggest that improvement in forecasters' discrimination skills (i.e., resolution) requires more than comprehension of the resolution concept and related performance feedback . . ." (p. 572).

The one exception to this general finding is provided by Sharp et al. (1988), who found an effect of personalized performance feedback on discrimination but not on calibration. However, this result can be explained by an important methodological difference between the research by Sharp et al. (1988) and that of Benson and Önkal (1992) and Lichtenstein and Fischhoff (1980).[2] In particular, Sharp et al. did not provide participants with a forced choice between two alternatives. Instead, participants were asked to generate the answers to general knowledge questions and to state a probability that they were correct, and all nonanswered questions (which totaled more than one third of the questions) were eliminated from the analysis, providing only a small number of judgments and thus a small number of judgment categories per participant. As discussed by Sharp et al. (1988, p. 280), the performance feedback may have increased evaluation apprehension, which in turn could have led to the use of more judgment categories. Further, as noted by Yaniv, Yates, and Smith (1991), the use of more forecast categories will increase one's discrimination score, even in the absence of any gain in true discriminatory ability. Thus, it seems plausible that the improvement in discrimination found by Sharp et al. was an artifact of the method employed in their experiment.

ENVIRONMENTAL FEEDBACK

Benson and Önkal (1992) suggest that environmental feedback, unlike performance feedback, should be effective for improving people's discrimination skill, since environmental information provides information about the event to be judged. Only a small amount of work, however, has examined the impact of environmental feedback isolated from other types of feedback on judgmental accuracy. Lichtenstein and Fischhoff (1977, Experiment 2) trained participants to discriminate between European and American handwriting by providing them with samples of each type of handwriting. This handwriting training served as a type of environmental feedback, as it provided the participants with task information. As predicted, those participants who underwent the training procedure achieved higher discrimination scores than did those who received no such training.

If calibration and discrimination are psychologically distinct concepts, then providing domain-specific information (environmental feedback) should have no impact on calibration. In fact, Lichtenstein and Fischhoff did find an improvement in calibration scores resulting from the training in their study. However, they concluded that this improvement did not reflect a true improvement in calibration skill, but instead resulted from the hard–easy effect (cf. Lichtenstein et al., 1982; Suantak, Bolger, & Ferrell, 1996), whereby difficult

[2] A second difference, and one we will return to later, is that Sharp et al. (1988) used a different measure of discrimination than did Lichtenstein and Fischhoff (1980). However, Benson and Önkal (1992) used the same measure of discrimination as did Sharp et al. (1988), so that cannot account for the difference between those two studies.

questions (those answered correctly 50–70% of the time) produce overconfidence, easy questions (those answered correctly 80–100% of the time) produce underconfidence, and those of moderate difficulty (those answered correctly 70–80% of the time) produce the best calibration. Since improvements in discrimination reflect gains in substantive knowledge on a topic, it would be expected that gains in discrimination would be accompanied by an increased number of questions answered correctly. Indeed, those participants who underwent the handwriting training answered 71% correctly while those who did not undergo the training answered only 51% correctly. Thus, on the basis of the increase in percentage of items answered correctly alone, the improvement in calibration could be attributed to the hard–easy effect rather than to a true improvement in calibration skill.

THE PRESENT STUDY

The previous review suggests that, within the domains studied, performance feedback improves calibration and that environmental feedback improves discrimination. There is also reason to believe that performance feedback does not affect discrimination and that environmental feedback does not affect calibration; however, these conclusions are more equivocal, in that past findings have been open to multiple interpretations. The primary goal of the present study, then, was to demonstrate this dissociation. To do this, we extended the previous research in two ways: first, we examined the effects of both performance and environmental feedback in the same domain, and, second, we attempted to overcome two potential artifactual explanations for the previous results.

Use of the Same Domain

One difficulty with comparing the results of previous studies that provided either performance or environmental feedback is that they used different tasks. It is conceivable that certain tasks lend themselves more readily to improvements in calibration or discrimination. Since those studies that have provided performance feedback have used different tasks (e.g., Benson & Önkal's (1992) participants judged football games) from those that provided environmental feedback (e.g., Lichtenstein & Fischhoff's (1977) participants judged whether handwriting was American or European), it is conceivable that the different results attributed to the type of feedback are instead caused by the task under investigation. To overcome this concern, the single domain of art history was used for all of our participants (control participants as well as those provided with either performance or environmental feedback). Specifically, all participants viewed slides of artwork and judged from which of two art history periods the artwork came. By so doing, we were able to examine the impact of both performance and environmental feedback within the same domain.

Elimination of Artifactual Explanations

Finally, our study was designed to overcome two potential artifactual explanations for previous results: the previously-discussed hard–easy effect and the measure of discrimination used. Recall that the hard–easy effect suggests that hard items (defined in terms of the percentage of items answered correctly) generally produce overconfidence, while easier items result in either a reduction of overconfidence or, occasionally, in underconfidence. To control for the hard–easy effect, then, we constructed two sets of slides, each with a different base rate of correct responses. In particular, we constructed one set of hard slides, which, according to the hard–easy effect, would be expected to produce overconfidence on the part of the participant. This situation is analogous to that used in the study by Lichtenstein and Fischhoff (1977). The second set of easy slides was designed to produce neither overconfidence nor underconfidence on the part of the participant. Thus, any gain in substantive ability (reflected in improved discrimination and percentage correct) not accompanied by a gain in calibration ability should, on the basis of the hard–easy effect alone, lead to reduced overconfidence with the hard slides but underconfidence with the easy slides. True gains in calibration ability, however, would lead to improvement with the hard slides without any detriment with the easy slides.

A second potential artifactual explanation pertains to the measure of discrimination used in the studies. Yates, Lee, Shinotsuka, Patalano, and Sieck (1998) discuss an important distinction between what they refer to as an external target event and an internal target event (see also Schneider, 1995; Yates, 1982). The example we discussed previously of judging the likelihood of a newborn being male (on a 0 to 100% scale) would be an example of an external target event, as the doctor is judging something external to him or her. Conversely, an internal target event would entail judging the likelihood of being correct, for example, by stating whether the newborn would be male or female and then judging the probability of being correct (usually, on a 50 to 100% scale). These situations are often treated as being mathematically equivalent, as it is assumed that a doctor who says there is an 80% chance that the infant will be female would say there is a 20% chance that the infant will be male. However, the discrimination measure is problematic for judgments of internal target events, as the concept of discrimination refers to being able to provide judgments that discriminate between the occurrence of two events, not between situations where one is right or wrong. As one example, a doctor who made judgments of 75 or 25% that an infant was male would demonstrate reasonable discrimination skill, as long as he or she was well-calibrated. However, if an internal target event were used, this doctor would always be making judgments of 75% of being correct and would be labeled as having no discrimination skill whatsoever. Thus Yates et al. (1998) recommend using an external target event whenever possible, which we did in the present study by having participants judge the probability that the artwork was from the later of two given time periods. Note that none of the previously-discussed studies on the effect of performance (Benson & Önkal, 1992; Lichtenstein & Fischhoff, 1980; Sharp

et al., 1988) or environmental (Lichtenstein & Fischhoff, 1977) feedback used an externally-defined target event.

A second advantage of using an external target event is that it overcomes a concern expressed by Sharp et al. (1988) and others (e.g., Benson & Önkal, 1992; Yaniv et al., 1991). Sharp et al. (1988) criticized the measure of discrimination used by Lichtenstein and Fischhoff (1980), as it is affected by the variability component of the mean probability score, $\overline{d}\,(1 - \overline{d})$, where $\overline{d}$ refers to the base rate of the target event. The interpretation of this variability component depends on whether the target event is internally or externally defined. In the case of an internal target event, the variability component is generally referred to as knowledge, as the score on this index varies according to the percentage of times the participant answers the item correctly. As Sharp et al. discuss, this situation is problematic, in that discrimination scores are affected by the knowledge component, which varies from participant to participant (as well as with training), and thus recommend that a correction be used. If, however, an external target event is used, then the variability component has a different meaning. In this situation, the variability component is affected solely by the proportion of times the target event occurs, which can be controlled easily in the experiment. Thus, by using an externally-defined target event, it is straightforward to hold the level of variability constant across participants, thereby making the discrimination scores comparable without employing the correction recommended by Sharp et al. Similarly, the concerns raised by Yaniv et al. (1991) are met by using an external target event as well as an equal number of response categories and trials for all participants.

METHOD

Participants

Participants in this study were 43 male and 41 female undergraduate students enrolled in introductory psychology, who participated as one method of fulfilling a course requirement. In order to obtain a trainable participant pool, all participants were not knowledgeable about art history. Specifically, students having completed any coursework in art history, having completed considerable coursework in the studio arts, or indicating considerable personal knowledge of art were ineligible to participate.

Design

The design was a 3×2 mixed factorial design, with type of training (none, performance feedback, or environmental feedback) as the between-subjects variable and stimulus difficulty (hard slides or easy slides) as the within-subjects variable.

Materials

Stimuli selection. This study required the use of approximately 200 slides of artwork, which were mostly of sculptures and paintings. In order to minimize

artifactual improvement in calibration skill with the acquisition of substantive knowledge due to the hard–easy effect, we required an equal number of easy and hard stimuli. Constructing these sets of slides required the administration of a pretest to determine which slide stimuli were easy and which were hard.

In this pretest, 360 slides of artwork were placed in a random order. These slides were drawn from the survey collection that accompanies Helen Gardner's *Art through the Ages* (1996). We selected 90 samples from each of the following periods of art history: Classical (500 BC–300 AD), Medieval (300–1400 AD), Renaissance (1400–1600 AD), and Impressionism (1800–1900 AD). Fifty students were then asked to choose between two given art history periods for each slide: one alternative was the correct period and the other was randomly chosen from among the remaining periods. In all cases choice A was chronologically the earlier of the two possible periods. In general, our participants did quite well at this task, with 79% of the answers being correct.

From this original pool of 360 slides, we chose 100 easy slides and 100 hard slides based on the performance of the participants in the pretest. As discussed previously, the hard–easy effect suggests that as the percentage correct increases, there will be a change from overconfidence to underconfidence. Suantak et al. (1996) identified the point that elicited neither overconfidence nor underconfidence to be about 77%. Hard items were chosen to produce overconfidence in untrained participants. In particular, slides that between 30 and 70% (M = 60%) of our participants answered correctly were selected as the hard stimuli. Easy items were chosen to produce neither overconfidence nor underconfidence in untrained participants. Specifically, slides that between 70 and 90% (M = 80%) of participants answered correctly were selected as easy stimuli. Thus, these slides were near to the absolute accuracy point of 77% identified by Suantak et al. (1996), but might be expected to produce mild underconfidence.

The final pool of 200 slides was then randomly organized into two separate presentations of 100 slides, 50 easy and 50 hard slides per presentation. The order in which the slides were presented was again determined randomly.

Response forms. For the actual study, participants were given the same two choices that were provided for that specific slide in the pretest. As with the pretest, choice A was always the earlier period and choice B the later period. Participants indicated the probability that the slide was from the later period by circling a probability judgment from 0–100%, given in increments of 10%.

Procedure

For each of six sessions, 14 or 15 participants gathered in a large room with a slide projector and viewing screen. Each session lasted roughly 2 hours, and participants were informed that the session would include three parts: pretraining, training, and posttraining. Participants were randomly assigned to the performance feedback, the environmental feedback, or the no feedback group, with the constraint that each group could consist of no more than five

participants per session. We ended with 25 participants in the no feedback group, 29 in the performance feedback group, and 30 in the environmental feedback group. Because of the nature of the training sessions, three experimenters were required to conduct the study.

Pretraining. The pretraining period lasted approximately 30 minutes. During the first ten minutes, Experimenter 1 read the instructions and provided a brief lecture on probability accuracy to all of the participants. Specifically, the experimenter informed participants that they would see slides of artwork drawn from four periods of art history and be asked to indicate the probability that the slide was drawn from the later of two given periods. It was particularly emphasized that judgments of 100% should mean that the participant was certain the slide was from the later time period, judgments of 0% that the participant was certain the slide was from the earlier period, and judgments of 50% that the participant was just guessing. Following this explanation of the task, participants were told that the three participants with the highest overall accuracy would receive a $20 cash prize. Further, participants were told that overall accuracy would be determined by two factors: discrimination and calibration. Given the complexity of the concept of discrimination, however, we did not actually define that term for the participants. Instead, we simply said that the first factor was related to their ability to determine whether the slide shown was from the earlier or the later period. Several minutes were then devoted to a basic discussion of calibration, during which time participants heard examples of and saw graphs (see Fig. 2) indicating perfect calibration, overconfidence, and underconfidence. It was emphasized that the probability rating circled on the response form should always reflect the likelihood that choice B is correct. For example, if the 80% probability category was circled ten times, then exactly eight of those ten slides ought to be from the later of the two given periods.

Experimenter 2 then presented the stimuli. Participants saw 100 slides of artwork (at a rate of five slides per minute) and made probability judgments

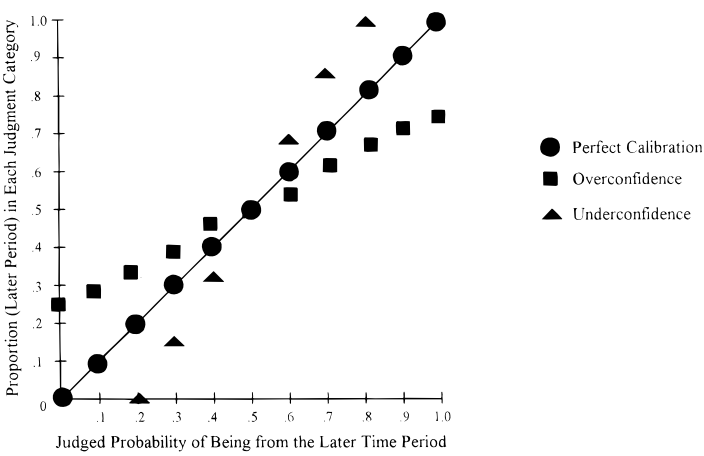

FIG. 2. Calibration graphs indicating perfect calibration, overconfidence, and underconfidence.

regarding the likelihood that the artwork was from the later of the two periods. Fifty hard and 50 easy slides were presented in random order.

Training. For the training session, which lasted approximately 45 minutes, participants were divided into the three experimental groups, each containing a maximum of five participants.

Experimenter 1 took the performance feedback group to another room, where they received personalized feedback regarding their calibration performance. Specifically, we constructed a calibration graph for the 100 pretest responses of each participant using Probability Analyzer (Yates, Purdy, & Potts, 1994). After each individual's calibration graph was produced, Experimenter 1 conducted a brief (usually 2–5 minutes) feedback session with each participant individually. This session consisted of an explanation of the calibration graph, as well as suggestions for improvement. In keeping with the advice of Lichtenstein et al. (1982), particular emphasis was placed on avoiding the use of extreme categories, when appropriate. Shortly before the end of this part of the experiment, Experimenter 1 asked each participant to briefly (usually 15–30 seconds) summarize the major advice given him or her for improvement. The vast majority of the participants correctly recalled the advice; the rest were given a brief additional training session. All participants were allowed to keep their calibration graph and notes to refer to during the posttraining test. To keep the participants occupied while they were not personally receiving training, they were given an unrelated task to perform during the 45-minute session as well.

The second group, the environmental feedback group, remained in the same room with Experimenter 2. Here they received environmental feedback in the form of a 30-minute lecture on art history, designed to increase their substantive knowledge about art history. Each participant was given a handout that described five to ten important characteristics of each of the four art history periods. Slides were shown and discussed to illustrate each characteristic. Participants took notes on their handouts, which they were allowed to use in the posttraining. Next, in order to facilitate the application of the knowledge acquired in the lecture, participants saw and were collectively quizzed on 20 slides from the pretraining. With the remaining time participants completed the same unrelated task completed by the performance feedback groups.

The control (no feedback) group was escorted by Experimenter 3 to a separate room, where they completed the same unrelated task that was completed by the performance feedback and environmental feedback groups.

Posttraining. The posttraining period lasted thirty minutes. All participants were reunited in the original room, and the same procedure used for the pretraining was repeated using a new set of slides. The same instructions were again read aloud by Experimenter 2; however, the more elaborate discussion of calibration using the calibration graph was omitted. Finally, the second series of 100 slides of artwork was presented at a rate of five slides per minute. Again, 50 hard slides and 50 easy slides were presented in random order.

RESULTS

Data were collected twice during the experiment, once during the pretraining period and once during the posttraining period. Table 1 presents the means for each of the conditions separately for both sessions. Although our primary interest was in discrimination and calibration (both overconfidence and calibration more generally), for completeness we report the mean probability score and percentage correct statistics as well (see below for how we determined percentage correct). It is worth noting that the only significant improvements in $\overline{PS}$ from pretraining to posttraining occurred in the performance feedback and environmental feedback conditions. This improvement would be expected because, as discussed previously, $\overline{PS}$ is just a composite of calibration and discrimination. Thus our expectation was that both types of feedback would improve $\overline{PS}$, but for different reasons.

Calibration and discrimination were measured by the calibration and discrimination indices that result from the partitioning of $\overline{PS}$ (see Murphy (1973) and Yates (1994) for details on the partitioning and formulas for the measures).

TABLE 1

Means of the Dependent Measures at Pretraining and Posttraining by Training Condition

	Training condition					
	Performance feedback		Environmental feedback		No feedback	
Dependent measure	Pre-training	Post-training	Pre-training	Post-training	Pre-training	Post-training
Hard stimuli						
Mean probability score[a]	.2931	.2593**	.2868	.2331**	.2862	.2729
Over/underconfidence[b]	.1877	.1059**	.1761	.1840	.1879	.1676
Calibration index[c]	.0947	.0673**	.0942	.0895	.0943	.0934
Discrimination index[d]	.0479	.0390	.0538	.0920**	.0545	.0561
Percent correct	56.2%	58.3%	57.0%	71.4%**	58.2%	60.4%
Easy stimuli						
Mean probability score[a]	.1874	.1757	.1770	.1520*	.1749	.1533
Over/underconfidence[b]	.0422	−.0220**	.0407	.1124**	.0494	.0228
Calibration index[c]	.0532	.0480	.0460	.0491	.0473	.0446
Discrimination index[d]	.1058	.1159	.1090	.1407**	.1124	.1349**
Percent correct	72.6%	74.6%	73.9%	81.1%**	74.8%	79.4%**

[a] Lower mean probability scores indicate better judgments.

[b] Negative over/underconfidence scores indicate underconfidence; positive scores indicate overconfidence.

[c] Lower calibration index scores indicate better calibration.

[d] Higher discrimination index scores indicate better discrimination.

* Paired *t* test comparing pretraining and posttraining means revealed significance at the .05 level.

** Paired *t* test comparing pretraining and posttraining means revealed significance at the .01 level.

Additionally, we calculated over/underconfidence for each of our participants as another measure of calibration. To do this, we first converted each of the participants' judgments so that they comprised an estimate of the correct answer and a corresponding probability of being correct, as if we had used an internal target event initially. For example, a judgment of an 80% chance that the artwork was from the later period was converted to an estimate that the artwork was from the later period, with a corresponding probability of .80. A judgment of .20 that the artwork was from the later period, however, was converted to an estimate that the artwork was from the earlier period, with a corresponding probability of .80. We then determined each participant's average probability rating of being correct ($\bar{f}$) and the corresponding percentage of correct answers ($\bar{d}$).[3] Finally, we computed the measure of over/underconfidence by subtracting the percentage of correct answers from the average probability judgment ($\bar{f} - \bar{d}$). It should be readily apparent, then, that positive values on this measure indicate overconfidence, while negative values indicate underconfidence. This approach for measuring over/underconfidence with an external target event has been used in a number of other studies (e.g., Yates et al., 1998).

We used the following analytic strategy for both the calibration and the discrimination measures. First, we examined whether there were changes from pretraining to posttraining in each of the different conditions. Paired *t* tests were used to compare pretraining scores with posttraining scores. Second, since there were some small changes even in the no feedback control group, we used that group as a baseline. Specifically, we calculated change scores from pretraining to posttraining on the relevant dependent measures and compared the change scores in both feedback groups to those of the control group by means of independent samples *t* tests. Table 2 lists these change scores, as well as which pairwise comparisons were significantly different at the .05 level. We chose this analytic approach rather than an omnibus one due to the fact that several of our primary hypotheses involved the lack of effects. Thus we felt it was particularly essential to maintain a high level of power for all the tests we conducted.

Finally, we constructed calibration plots for the posttraining judgments for the performance feedback group, the environmental feedback group, and the no feedback group (see Figs. 3, 4, and 5, respectively). It is worth emphasizing that these diagrams are aggregated over all the judgments of all the participants in the particular group and thus that the judgments that comprise them are not all independent. Nonetheless, they provide a useful means for determining precisely why any differences between the groups occurred.

[3] The only complication arose when participants estimated that there was a 50% chance that the artwork was from the later time period. In this situation, we kept the probability estimate of .50 and used .50 for *d*. Thus, these situations had no impact on the over/underconfidence score.

TABLE 2

Dependent Measure Change Scores (Posttraining Scores Minus Pretraining Scores) by Training Condition

	Training condition		
Dependent measure	Performance feedback	Environmental feedback	No feedback
Hard stimuli			
Mean probability score[a]	$-.0392_{xy}$	$-.0537_{x}$	$-.0133_{y}$
Over/underconfidence[b]	$-.0818_{x}$	$.0079_{y}$	$-.0203_{y}$
Calibration index[c]	$-.0247_{x}$	$-.0047_{x}$	$-.0009_{x}$
Discrimination index[d]	$-.0089_{x}$	$.0382_{y}$	$.0016_{x}$
Percent correct[e]	$.0217_{x}$	$.1440_{y}$	$.0220_{x}$
Easy Stimuli			
Mean probability score[a]	$-.0117_{x}$	$-.0250_{x}$	$-.0216_{x}$
Over/underconfidence[b]	$-.0642_{x}$	$.0717_{y}$	$-.0266_{x}$
Calibration index[c]	$-.0052_{x}$	$.0031_{x}$	$-.0027_{x}$
Discrimination index[d]	$.0101_{x}$	$.0317_{x}$	$.0225_{x}$
Percent correct[e]	$.0203_{x}$	$.0717_{y}$	$.0452_{xy}$

Note. Dependent measure change scores with different subscripts within the same row are different at the .05 level of significance, as shown by independent samples t-tests.

[a] Negative mean probability change scores indicate improvement from pretraining to posttraining.

[b] Negative over/underconfidence change scores indicate improvement (reduced overconfidence) from pretraining to posttraining.

[c] Negative calibration index change scores indicate improvement from pretraining to posttraining.

[d] Positive discrimination index change scores indicate improvement from pretraining to posttraining.

[e] Positive percent correct change scores indicate improvement from pretraining to posttraining.

Calibration Measures

As discussed previously, we examined two measures of calibration: overconfidence and the calibration index. Although overconfidence is just one type of miscalibration, we are focusing our analysis on the overconfidence measure (rather than on the calibration index) for both theoretical and empirical reasons. First, there have recently been a number of concerns expressed about how to interpret calibration (see, e.g., Dawes & Mulford, 1996; Erev, Wallsten, & Budescu, 1994). Although the issues are complex, the crux of the concern is that there is some error in the judgment process or, relatedly, an imperfect relationship between confidence and accuracy. This results in the calibration curve being flatter than the line of perfect calibration, resulting in apparent (but not necessarily actual, according to these authors) overconfidence. Another way to see the problem is that the measure of calibration conditions on the judgment made, i.e., it examines the percentage of correct answers *given* the probability judgment category (Parker, Downs, Fischhoff, Bruine de Bruin, & Dawes, 1999). Importantly, the measure of over/underconfidence that we are

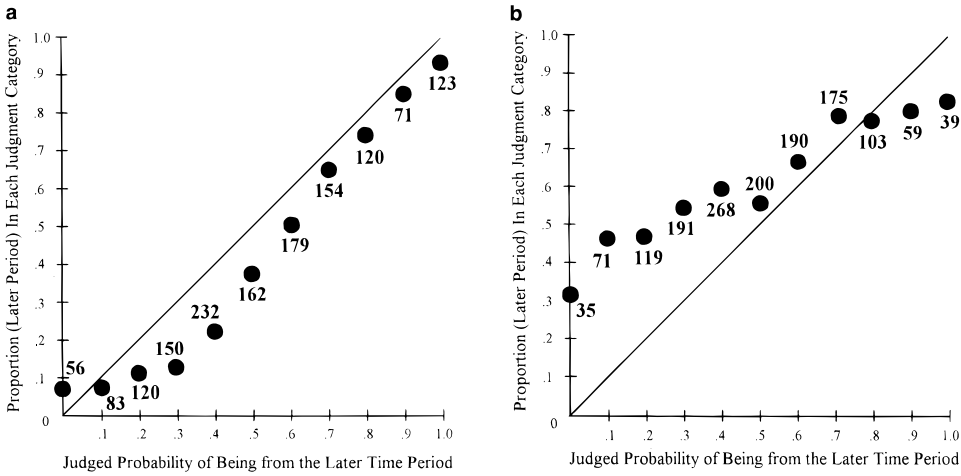

FIG. 3. Posttraining calibration graphs for the performance feedback group: (a) Easy slides; (b) hard slides. Note that the numbers by each of the points indicate the number of times that judgment category was used aggregated over all the participants in the performance feedback group.

using does not condition on either the judgment made or the outcome; thus, this measure is not affected by this concern. (See also Brenner, Koehler, Liberman, & Tversky, 1996.)

Second, although there are other types of miscalibration (such as overpredicting the presence of an externally-defined target event), by far the most frequently discussed form of miscalibration is over- or (less frequently) underconfidence (see, e.g., Plous (1993); see also Winman & Juslin (1993) for some empirical evidence that over/underconfidence is the primary contributor to miscalibration). Thus, the results with the calibration index could reasonably

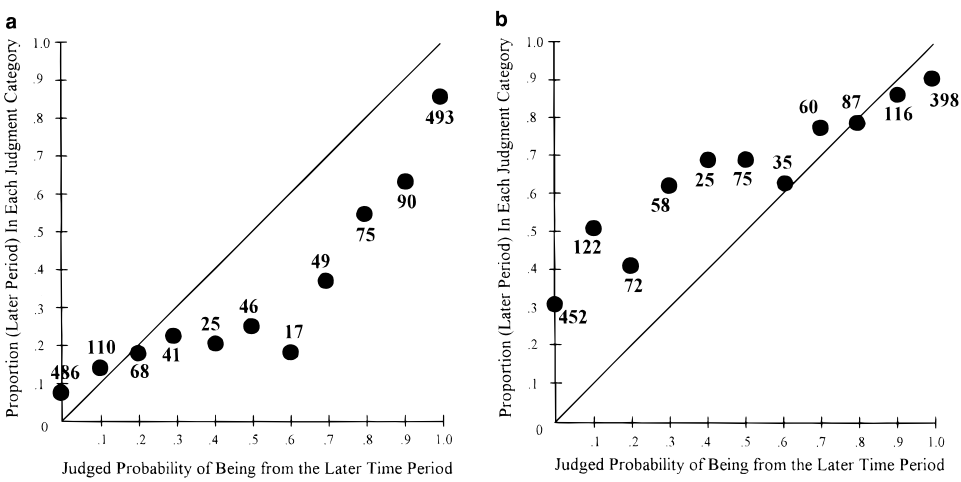

FIG. 4. Posttraining calibration graphs for the environmental feedback group: (a) Easy slides; (b) hard slides. Note that the numbers by each of the points indicate the number of times that judgment category was used aggregated over all the participants in the environmental feedback group.

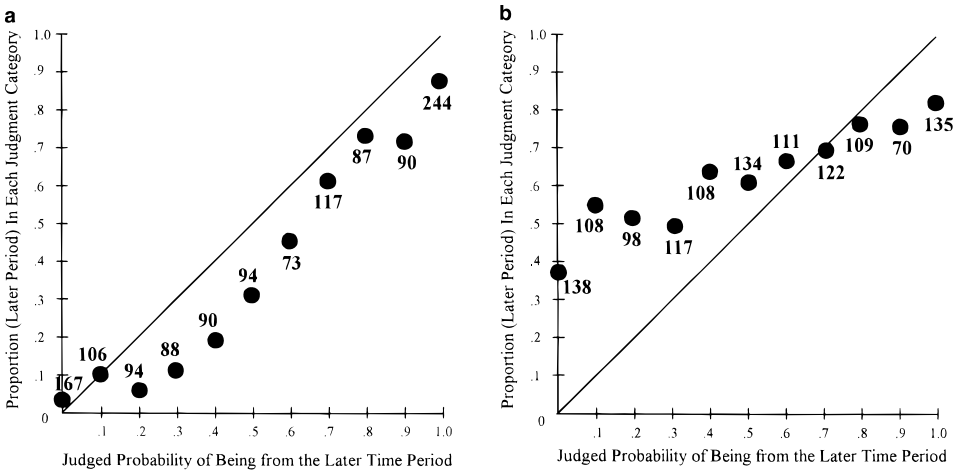

FIG. 5. Posttraining calibration graphs for the no feedback group: (a) Easy slides; (b) hard slides. Note that the numbers by each of the points indicate the number of times that judgment category was used aggregated over all the participants in the no feedback group.

be expected to be weaker than for overconfidence, as the effect of overconfidence would be diluted by the other (generally random) factors that could potentially contribute to miscalibration. In support of this line of reasoning, at least one recent study (Bornstein & Zickafoose, 1999) found effects on overconfidence but not on calibration more generally. Thus, for the above reasons and for ease of reporting, we report our results primarily in terms of the over/underconfidence measure, but note where the calibration results are qualitatively different.

As shown in Table 1, considerable overconfidence was found for the hard items, but there was still a small amount of overconfidence for the easy items, as opposed to what we predicted from the hard–easy effect. The amount of overconfidence for the easy items was small, however, and can be easily explained by the fact that participants on average did not quite answer 77% of the items correctly, as well as by the fact that the crossover point from over- to underconfidence varies slightly according to the particular domain. As predicted, overconfidence decreased as a result of performance feedback for both hard and easy slides, $t(28) = 3.919$, $p = 001$, and $t(28) = 2.902$, $p = .007$, respectively (see Table 1). For easy slides, this improvement resulted in mild underconfidence in posttraining, $M = -.022$. Participants in the environmental feedback group actually displayed worsened overconfidence from pretraining to posttraining, though it only reached significance for the easy stimuli, $t(29) = 5.295$, $p < .001$. The no feedback group demonstrated only nominal overconfidence improvements for both hard and easy stimuli (both p's $> .15$).[4]

Table 2 shows the change scores from pretraining to posttraining. For hard stimuli, the decrease in overconfidence for the performance feedback group

[4] The results with the calibration index were qualitatively similar for the hard items. However, for the easy items, although the trends were in the same direction, the improvement under performance feedback and the reduced accuracy under environmental feedback were both nonsignificant.

was significantly greater than for the no feedback control group, $t(52) = 1.998$, $p = .05$. For the easy slides, the same pattern of results was found, but it did not reach significance, $t(52) = 1.274$, $p = .21$. For neither set of slides did the environmental feedback group produce a greater reduction in overconfidence than the control group, and it actually produced a significantly greater increase in overconfidence for the easy items, $t(53) = 4.324$, $p < .001$.[5]

To see why the improvements in overconfidence occurred, consider the calibration graphs shown in Figs. 3–5. In terms of the hard slides, there are two trends evident for all three groups. First, there was a tendency to underestimate the presence of the target event, as indicated by the points generally being to the left of the line of perfect calibration. This indicates that the hard slides used in this study were believed to be from the earlier periods more than was actually the case, perhaps due to participants believing that the reason the particular artwork was difficult to classify was because it was from a long time ago. (Analogous reasoning can be used to explain the overestimation of the target event for the easy slides.) Second, there was general overconfidence, as indicated by the fact that a line connecting the points on the calibration graph would be too horizontal. Indeed, although the points for the performance feedback group are slightly closer to the line of perfect calibration than for the other groups, the differences are quite small.

An examination of the number of judgments that comprise each data point, however, tells a different story. Note that—for both the easy and the hard slides—the most frequently used judgment categories by participants in both the environmental feedback and the control group were the two extreme categories of 0.0 and 1.0, and that these judgment categories were generally the worst calibrated, especially for the hard slides. Participants in the performance feedback group, however, rarely used these categories, with most of their judgments being between .3 and .7. Thus the relatively infrequent use of the extreme judgment categories by participants in the performance feedback group was primarily responsible for their reduced overconfidence.

Discrimination

As shown in Table 1, environmental feedback improved scores on the discrimination index for both hard slides, $t(29) = 5.368$, $p < .001$, and easy slides, $t(29) = 3.364$, $p = .002$. Conversely, performance feedback had no effect on discrimination for either hard or easy items (both p's $> .15$). Surprisingly, in the no feedback group there was an improvement from pretraining to posttraining for the easy items, $t(24) = 3.055$, $p = .005$, though this effect was not found for the hard items, $t(24) = .295$, $p = .77$.

As shown in Table 2, the improvement in discrimination was greater for the

[5] The difference between the performance feedback and control groups for the hard items, though similar in magnitude for over/underconfidence and the calibration index, did not quite reach significance for the calibration index, $t(52) = 1.856$, $p = .07$. Additionally, there were no differences between the environmental feedback and control groups on the calibration index for either set of stimuli.

environmental feedback group than for the no feedback control group, although this difference only reached significance for the hard items ($t(53) = 3.944$, $p < .001$, for hard items, and $t(53) = .743$, $p = .46$, for easy items). As indicated above, the failure of the environmental feedback group to outperform the control group for the easy items is probably due to the unexpected improvement in discrimination on easy items for the control participants. Most importantly, there was no evidence that the performance feedback group demonstrated greater improvement in discrimination than did the control group. In fact, for both hard and easy items, the improvement was greater for the control group than for the performance feedback group, although neither difference approached significance (both p's $> .20$).

As with the calibration results, the improvements in discrimination are reflected by the number of participants choosing each of the judgment categories. Although simply making more extreme judgments will not improve discrimination (see Yates, 1990), more extreme judgments accompanied by an increase in knowledge does improve discrimination scores. That the participants in the environmental feedback group did have increased knowledge is indicated by the fact that, although they had many more extreme judgments than did participants in the other two groups, these extreme judgments were correct approximately as often as were the extreme judgments of the other two groups. This finding would not be possible if the environmental feedback group had not gained any substantive knowledge as part of their training.

DISCUSSION

Recall that the primary goal of this study was to demonstrate a dissociation, in that we expected performance feedback to improve calibration but not discrimination and environmental feedback to improve discrimination but not calibration. The results strongly supported this hypothesis. Additionally, we found two unexpected effects: (1) the impact of feedback was greater for hard slides than for easy slides, and (2) environmental feedback led to increased overconfidence for easy slides. We shall proceed as follows. First, we discuss the effect of performance feedback on calibration and environmental feedback on discrimination, focusing on why the effects were stronger for hard slides than for easy slides. Second, we discuss the implications of the lack of effect of performance feedback on discrimination and environmental feedback on calibration. Third, we discuss the conditions under which one might expect environmental feedback to affect calibration measures, focusing on why environmental feedback led to increased overconfidence in our work but not in previous research (e.g., Lichtenstein & Fischhoff, 1977). And finally, we discuss some ways in which our results might reasonably be extended.

Differential Effects on the Hard and Easy Slides

In retrospect, it is not surprising that there were stronger effects with the hard slides than with the easy slides, as there was more room for improvement

with the hard items in terms of both discrimination and overconfidence. Discrimination in the absence of training was much better for the easy items (M = .11) than for the hard items (M = .05). Thus, there was more room to improve with the hard slides than with the easy slides. Indeed, much of the environmental feedback involved cues that were intended to be useful for some of the more difficult discriminations. This can be seen as well in the fact that with environmental feedback an additional 14.4% of the hard items were answered correctly, but only an additional 7.2% of the easy items were. Thus, the greater impact of environmental feedback on discrimination for the hard items is a relatively uninteresting result, as it simply indicates that it is easier to improve mediocre performance than performance that is already good.

The results regarding overconfidence are similar. Recall that we included easy slides to differentiate true gains in calibration ability resulting from environmental feedback from artifactual improvements due to the hard–easy effect under the assumption that participants would be neither over- nor underconfident for the easy slides. This inclusion turned out to be unnecessary, as, contrary to the results of Lichtenstein and Fischhoff (1977), the increase in items answered correctly by participants in the environmental feedback group was not associated with reduced overconfidence, an issue we will return to shortly. For the present purposes, the greater impact of performance feedback on the hard slides than on the easy slides again simply illustrates the fact that there was more room for improvement with the hard slides.

Independence of Calibration and Discrimination

Most importantly, there was no evidence that providing performance feedback led to improvements in discrimination or that providing environmental feedback led to improvements in calibration. We take this as strong evidence that calibration and discrimination reflect psychologically distinct skills. Stone and Hoffman (1999) have referred to these skills as *calibration expertise* and *substantive expertise*, respectively. Substantive expertise refers to domain-specific knowledge in a certain area, while calibration expertise reflects the ability to assign well-calibrated probability judgments on the basis of that knowledge (see also Ayton, 1992; Benson, Curley, & Smith, 1995; Bornstein & Zickafoose, 1999; Keren, 1991; Lichtenstein et al., 1982; Shanteau, 1988; Winkler & Murphy, 1968).

Despite the evidence for the independence of these two types of expertise, most formalized education assumes, at least implicitly, that the abilities are related. For example, most medical and clinical psychology programs provide substantive training in the field (for example, teaching indicators of types of illness), but (with some notable exceptions) provide little in the way of guidance as to how to translate this knowledge into well-calibrated probability judgments. This relative focus on substantive skills rather than on calibration skills may in part explain why experts in these domains frequently perform poorly on calibration tasks. Conversely, fields that do provide calibration training (e.g., meteorology) generally produce experts who are much better calibrated

(see, e.g., Ayton, 1992; Lichtenstein et al., 1982; Murphy & Winkler, 1984; Wallsten & Budescu, 1983).

In contrast to this general trend, however, there have been some recent educational efforts that have focused solely on calibration skills. As one example, Smith and Dumont (1997) describe a training program designed in particular to reduce overconfidence on the part of their clinical trainees. A major part of this regimen involves providing performance feedback in situations where the trainees are overconfident. The present work suggests that this type of training procedure is essential to the development of good calibration skills.

Impact of Substantive Training on Overconfidence

Additionally, our work suggests that, in the absence of appropriate calibration training, substantive training in the form of environmental feedback can actually *lower* calibration by increasing overconfidence. Recall that, after participants received environmental feedback, overconfidence increased. This was especially true with the easy slides, but there was even a small trend to this effect with the hard slides, where a substantial reduction in overconfidence was expected due to the hard–easy effect.[6] These results mirror a well-known study by Oskamp (1965). In his study, he provided information to clinical psychologists and psychology students about a 29-year-old man named Joseph Kidd. He then asked the participants a number of difficult questions about Joseph Kidd and asked them to provide probability judgments that their answers were correct. When the participants were provided only a small amount of information, their probability judgments were roughly in line with the actual percentage of items answered correctly. However, as more information was presented, their accuracy did not increase substantially, but their confidence in their judgments did, leading to greater amounts of overconfidence.

The difference between the study by Oskamp and the present work is that in our study, the information provided to participants *was* useful, as indicated by their improvements in discrimination and percentage correct. Thus, our study better reflects a training procedure where diagnostic information is being conveyed. Nonetheless, the increase in accuracy as measured by discrimination and percentage correct was accompanied by an even greater increase in confidence, leading to increased overconfidence. Note this result is in direct conflict with the findings of Lichtenstein and Fischhoff (1977), who found that with substantive training overconfidence decreased.

A clue to the cause of this difference can be seen in the research paradigm where participants are provided with outcome feedback, i.e., information on the outcome of different trials, but no explicit performance or environmental feedback. Two studies that provided outcome feedback are particularly relevant (see also Benson & Önkal, 1992; Einhorn, 1982; Fischer, 1982; Pulford & Colman, 1997; Subbotin, 1996). Arkes, Christensen, Lai, and Blumer (1987)

[6] It is worth emphasizing that these results do not in any sense contradict the hard–easy effect. In fact, given the results with the easy items, it seems likely that the hard–easy effect was responsible for the relative lack of effect with the hard items.

provided subjects with five practice questions, all of which appeared easy but were in fact quite difficult, and gave outcome feedback to half of the subjects. While no-feedback subjects were overconfident on 30 subsequent items, the outcome feedback subjects showed a slight trend toward underconfidence. Conversely, Stankov and Crawford (1997) found that providing outcome feedback on a vocabulary test increased the level of overconfidence. Why might outcome feedback have such different effects in the two studies? Arkes et al. suggest that their manipulation was successful because the participants were sufficiently surprised by how poorly they did on the practice questions that they lowered their confidence estimates later. In Stankov and Crawford's study, however, participants raised their confidence levels, presumably because they felt they were performing better than they were. This suggests a general rule, which is that the impact of feedback on overconfidence is highly dependent on the extent it conveys to the participant an accurate indication of how they are performing (thus essentially serving as performance feedback). In keeping with this conjecture, Petrusic and Baranski (1997) found that outcome feedback improved calibration in a difficult context (where the feedback would provide information as to how poorly the participants were doing) but not in an easy context.

This same explanation can explain the differences between the results of our research and those of Lichtenstein and Fischhoff (1977). Recall that as part of our environmental feedback, we listed a number of important characteristics of each of the art history periods. These were conveyed in a relatively absolutist fashion, and participants may well have expected that if they learned these characteristics they would be able to answer the vast majority of the questions correctly. In other words, the training procedure may have led them to think they knew more than they did. Conversely, the training given by Lichtenstein and Fischhoff involved the participants' studying the test stimuli themselves, and participants needed to construct any rules on their own. Thus, it is reasonable that participants in their study did not expect to improve as much as those in our study did. Indeed, it is worth noting that the behavior of our participants can be considered to be completely rational. They knew that they had learned more information, and without any firm basis to judge how much more they had learned, they made the reasonable (but incorrect) assumption that they had learned quite a bit, which led to the increased overconfidence. Again, the point we wish to emphasize is that without performance feedback, it is extremely difficult to make well-calibrated probability judgments. Our data suggest there is little reason to believe that other types of training, such as providing participants with environmental feedback, should be useful for helping people make better calibrated judgments.

Future Directions

Although our research investigated probabilistic judgment, it is informative to speculate briefly on how these results relate to the literature on deterministic judgment. In their review of the literature on cognitive feedback, Balzer, Doherty, and O'Connor (1989) outlined three types of cognitive feedback that could

be expected to improve participants' deterministic judgments: task information (information about the task to be judged), cognitive information (information about the participant's judgments), and functional validity information (information about the accuracy of the participant's judgments). It should be evident that task information can be considered a form of environmental feedback, and functional validity information a form of performance feedback. Given our results, then, it would be reasonable to expect that the types of feedback (at least task information and functional validity information) identified by Balzer et al. would have separate effects on different measures of judgmental accuracy.

Along these lines, recent work conducted by Balzer and his colleagues (Balzer et al., 1994; Balzer, Sulsky, Hammer, & Sumner, 1992; see also Remus, O'Connor, & Griggs, 1996) found that task information was the primary contributor to improvements due to cognitive feedback in multiple cue probability learning tasks. Specifically, neither cognitive information nor functional validity information improved participants' accuracy beyond that provided by task information. Note, however, that the dependent measures used in these tasks assessed the participant's ability to make point judgments that were correlated with the actual outcomes; in none of these studies was participants' estimates of uncertainty examined. Our results suggest that if the procedure employed by Balzer and colleagues was extended to ask participants about the level of uncertainty associated with their judgments, then other types of feedback (functional validity information in particular) could reasonably be expected to be successful. Indeed, research using the fractile method (see Alpert & Raiffa, 1969) to assess participants' uncertainty has found that individualized performance feedback does reduce overconfidence (cf. Roth, 1993). More work specifically examining the impact of all three types of cognitive feedback on uncertainty estimation in multiple cue probability learning tasks would be quite valuable.

Finally, it is worth emphasizing that although our research demonstrated that substantive and calibration expertise develop through different routes, it did not demonstrate precisely what those routes are. Nonetheless, the present results are in keeping with existing process models. In particular, there is increasing evidence that the formation of judgments and responses based on those judgments occur in distinct stages (e.g., Wallsten et al., 1999). It is plausible that substantive expertise is related to the formation of the judgments while calibration expertise involves the translation of those judgments into overt responses (see also Ferrell & McGoey, 1980; Suantak et al., 1996), which could in part explain the dissociation found in the present research. Future research should systematically examine this possibility.

Conclusion

The purpose of this paper was to determine how training should be conducted to improve the accuracy of people's judgments. Our answer to this question is that different types of training are required to improve different aspects of probability judgment. On the basis of this research, environmental feedback

appears to be necessary for improvements in discrimination and other measures of substantive expertise, while performance feedback is required for improvements in calibration. Moreover, without accompanying performance feedback, substantive training can actually lead to greater overconfidence. We suggest, then, that it is essential to include both types of feedback in training programs.

REFERENCES

Adams, P. A., & Adams, J. K. (1958). Training in confidence judgments. *American Journal of Psychology*, **71,** 747–751.

Alpert, M., & Raiffa, H. (1969). *A progress report on the training of probability assessors*. Unpublished manuscript.

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

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

Balzer, W. K., Doherty, M. E., & O'Connor, R., Jr. (1989). Effects of cognitive feedback on performance. *Psychological Bulletin*, **106,** 410–433.

Balzer, W. K., Hammer, L. B., Sumner, K. E., Birchenough, T. R., Martens, S. P., & Raymark, P. H. (1994). Effects of cognitive feedback components, display format, and elaboration on performance. *Organizational Behavior and Human Decision Processes*, **58,** 369–385.

Balzer, W. K., Sulsky, L. M., Hammer, L. B., & Sumner, K. E. (1992). Task information, cognitive information, or functional validity information: Which components of cognitive feedback affect performance? *Organizational Behavior and Human Decision Processes*, **53,** 35–54.

Benson, P. G., Curley, S. P., & Smith, G. F. (1995). Belief assessment: An underdeveloped phase of probability elicitation. *Management Science*, **41,** 1639–1653.

Benson, P. G., & Önkal, D. (1992). The effects of feedback and training on the performance of probability forecasters. *International Journal of Forecasting*, **8,** 559–573.

Bolger, F., & Wright, G. (1992). Reliability and validity in expert judgment. In G. Wright and F. Bolger (Eds.), *Expertise and decision support* (pp. 47–76). New York: Plenum Press.

Bornstein, B. H., & Zickafoose, D. J. (1999). "I know I know it, I know I saw it": The stability of the confidence-accuracy relationship across domains. *Journal of Experimental Psychology: Applied*, **5,** 76–88.

Braun, P. A., & Yaniv, I. (1992). A case study of expert judgment: Economists' probabilities versus base-rate model forecasts. *Journal of Behavioral Decision Making*, **5,** 217–231.

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

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

Chan, S. (1982). Expert judgments under uncertainty: Some evidence and suggestions. *Social Science Quarterly*, **63,** 428–444.

Christensen-Szalanski, J. J. J., & Bushyhead, J. B. (1981). Physicians' use of probabilistic information in a real clinical setting. *Journal of Experimental Psychology: Human Perception and Performance*, **7,** 928–935.

Dawes, R. M., & Mulford, M. (1996). The false consensus effect and overconfidence: Flaws in judgment or flaws in how we study judgment? *Organizational Behavior and Human Decision Processes*, **65,** 201–211.

Eddy, D. M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 249–267). Cambridge, UK: Cambridge University Press.

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

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

Ferrell, W. R., & McGoey, P. J. (1980). A model of calibration for subjective probabilities. *Organizational Behavior and Human Decision Performance*, **26,** 32–53.

Fischer, G. W. (1982). Scoring-rule feedback and the overconfidence syndrome in subjective probability forecasting. *Organizational Behavior and Human Performance*, **29,** 352–369.

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

Garb, H. N. (1989). Clinical judgment, clinical training, and professional experience. *Psychological Bulletin*, **105,** 387–396.

Gardner, H. (1996). *Gardner's art through the ages* (R. G. Tansey & F. S. Kleiner, Eds., 10th Ed.). New York: Harcourt Brace College.

Kahneman, D., & Lovallo, D. (1993). Timid choices and bold forecasts: A cognitive perspective on risk taking. *Management Science*, **39,** 17–31.

Keren, G. (1987). Facing uncertainty in the game of bridge: A calibration study. *Organizational Behavior and Human Decision Processes*, **39,** 98–114.

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

Liberman, V., & Tversky, A. (1993). On the evaluation of probability judgments: Calibration, resolution, and monotonicity. *Psychological Bulletin*, **114,** 162–173.

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). Cambridge, UK: Cambridge University Press.

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

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

Murphy, A. H., & Winkler, R. L. (1984). Probability forecasting in meteorology. *Journal of the American Statistical Association*, **79,** 489–500.

Oskamp, S. (1962). The relationship of clinical experience and training methods to several criteria of clinical prediction. *Psychological Monographs*, **76** (28, Whole No. 547), 1–25.

Oskamp, S. (1965). Overconfidence in case-study judgments. *Journal of Consulting Psychology*, **29,** 261–265.

Parker, A. M., Downs, J. S., Fischhoff B., Bruine de Bruin, W., & Dawes, R. M. (1999). *The appropriateness of adolescents' confidence in their knowledge: AIDS-related and general*. Manuscript submitted for publication.

Payne, D. G., & Wenger, J. J. (1998). *Cognitive psychology.* Boston: Houghton Mifflin.

Petrusic, W. M., & Baranski, J. V. (1997). Context, feedback, and the calibration and resolution of confidence in perceptual judgments. *American Journal of Psychology*, **110,** 543–572.

Plous, S. (1993). *The psychology of judgment and decision making*. New York: McGraw-Hill.

Pulford, B. D., & Colman, A. M. (1997). Overconfidence: Feedback and item difficulty effects. *Personality and Individual Differences*, **23,** 125–133.

Remus, W., O'Connor, M., & Griggs, K. (1996). Does feedback improve the accuracy of recurrent judgmental forecasts? *Organizational Behavior and Human Decision Processes*, **66,** 22–30.

Roth, P. L. (1993). Research trends in judgment and their implications for the Schmidt-Hunter global estimation procedure. *Organizational Behavior and Human Decision Processes*, **54,** 299–319.

Schneider, S. (1995). Item difficulty, discrimination, and the confidence-frequency effect in a categorical judgment task. *Organizational Behavior and Human Decision Processes*, **61,** 148–167.

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

Sharp, G. L., Cutler, B. L., & Penrod, S. D. (1988). Performance feedback improves the resolution of confidence judgments. *Organizational Behavior and Human Decision Processes*, **42,** 271–283.

Smith, D., & Dumont, F. (1997). Eliminating overconfidence in psychodiagnosis: Strategies for training and practice. *Clinical Psychology: Science and Practice*, **4,** 335–345.

Stankov, L., & Crawford, J. D. (1997). Self-confidence and performance on tests of cognitive abilities. *Intelligence*, **25,** 93–109.

Stone, E. R., & Hoffman, R. R. (1999). *The distinction between calibration and substantive expertise in the evaluation of judgment tasks*. Unpublished manuscript.

Suantak, L., Bolger, F., & Ferrell, W. R. (1996). The hard-easy effect in subjective probability calibration. *Organizational Behavior and Human Decision Processes*, **67,** 201–221.

Subbotin, V. (1996). Outcome feedback effects on under- and overconfident judgments (general knowledge tasks). *Organizational Behavior and Human Decision Processes*, **66,** 268–276.

Wallace, H. A. (1923). What is in the corn judge's mind? *Journal of the American Society of Agronomy*, **15,** 300–304.

Wallsten, T. S., Bender, R. H., & Li, Y. (1999). Dissociating judgment from response processes in statement verification: The effects of experience on each component. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **25,** 96–115.

Wallsten, T. S., & Budescu, D. V. (1983). Encoding subjective probabilities: A psychological and psychometric review. *Management Science*, **29,** 151–173.

Wallsten, T. S., & González-Vallejo, C. (1994). Statement verification: A stochastic model of judgment and response. *Psychological Review*, **101,** 490–504.

Wigton, R. S. (1988). Applications of judgment analysis and cognitive feedback to medicine. In B. Brehmer & R. B. Joyce (Eds.), *Human judgment: The SJT approach*. Amsterdam: North-Holland.

Winkler, R. L., & Murphy, A. H. (1968). "Good" probability assessors. *Journal of Applied Meteorology*, **7,** 751–758.

Winman, A., & Juslin, P. (1993). Calibration of sensory and cognitive judgments: Two different accounts. *Scandinavian Journal of Psychology*, **34,** 135–148.

Yaniv, I., Yates, J. F., & Smith, J. E. K. (1991). Measures of discrimination skill in probabilistic judgment. *Psychological Bulletin*, **110,** 611–617.

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

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

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

Yates, J. F., Lee, J.-W., Shinotsuka, H., Patalano, A. L., & Sieck, W. R. (1998). Cross-cultural variations in probability judgment accuracy: Beyond general knowledge overconfidence? *Organizational Behavior and Human Decision Processes*, **74,** 89–117.

Yates, J. F., Purdy, G. N., & Potts, P. R. (1994). *Probability analyzer* [computer program]. Ann Arbor: Department of Psychology & Office of Instructional Technology, University of Michigan.

Yates, J. F., Zhu, Y., Ronis, D. L., Wang, D.-F., Shinotsuka, H., & Toda, M. (1989). Probability judgment accuracy: China, Japan, and the United States. *Organizational Behavior and Human Decision Processes*, **43,** 147–171.

Received January 14, 2000; published online September 14, 2000