Journal of Experimental Psychology:
Learning, Memory, and Cognition
1994, Vol. 20, No. 2, 461-469

0278-7393/94/$3.00

Hypothesis Generation and Confidence in Judgment

Derek J. Koehler

Ss asked to generate their own hypotheses expressed less confidence that they were true than did other Ss who were presented with the same hypotheses for evaluation. This finding holds across domains varying from prediction and social inference to general knowledge questions. Furthermore, Ss who generated their own hypotheses appeared to be more sensitive to their accuracy than were Ss who evaluated the hypotheses. The results are interpreted as evidence that the hypothesis generation task leads Ss to consider more alternative hypotheses than Ss who are asked to evaluate a prespecified hypothesis. This interpretation is supported by experiments demonstrating that the difference between generation and evaluation disappears if a closed set of alternatives is specified or if a delay is inserted between hypothesis generation and confidence assessment.

Although we believe a great many things, we hold some of our beliefs with greater conviction than others. Likewise, we express varying degrees of confidence in judgments we make under conditions of uncertainty. Ideally, if our sense of confidence is a reliable indicator of the accuracy of our judgments, we can use it to determine the weight to give some judgment when making a decision. For this and other reasons, researchers studying human judgment and social cognition have turned their attention increasingly to the problem of confidence in judgment.

In a domain in which accuracy can be measured, subjects are asked to make judgments and to estimate the probability that each answer they give is correct. The advantage of this approach is that it allows researchers to compare the estimated probability of being correct (i.e., confidence) with the actual percent correct. If people were perfectly calibrated (i.e., if their confidence estimates correctly reflected their accuracy), then about 50% of all the judgments that were assigned a 50% confidence estimate should be correct. Likewise, about 60% of judgments assigned a 60% confidence estimate should be correct, and so forth. At the very least, mean confidence should approximately equal mean accuracy.

Instead, the overwhelming finding is that people's confidence estimates tend to be too high, at least for tasks that offer a moderate to high level of difficulty. Researchers interested in this phenomenon have accumulated a substantial body of evidence demonstrating people's overconfidence in the truth of their beliefs and in the accuracy of their judgments across a wide variety of domains (e.g., Lichtenstein & Fischhoff, 1977; Lichtenstein, Fischhoff, & Phillips, 1982; Oskamp, 1965). Although such demonstrations are often reported in the literature, there has been significantly less research conducted to identify the sources of overconfidence. Obviously, because such a robust phenomenon is almost certainly overdetermined and partially context dependent, there is little chance of producing a single, well-specified psychological model that adequately accounts for overconfidence. Nevertheless, the direction of recent research (e.g., Griffin & Tversky, 1992; Peterson & Pitz, 1988) is toward identifying at least some of the major determinants or components of overconfidence.

The one idea that has captured attention in this area since the beginning is that overconfidence arises because viable alternatives to the focal hypothesis are often neglected. From this view, people are overconfident in their judgments because they focus exclusively on a preferred hypothesis and do not devote enough resources to considering other possibilities that might instead be true.

The most frequently cited evidence supporting this contention comes from a study by Koriat, Lichtenstein, and Fischhoff (1980). Subjects who were asked to answer two-alternative general knowledge questions were better calibrated (both because they were less confident and because their answers were more accurate) if they first wrote down all the reasons they could think of supporting and opposing each alternative before circling their choice and giving their probability estimate. A second experiment provided some evidence suggesting that most of this effect was attributable to the process of writing reasons that contradicted or opposed one's initial answer. Hoch (1985) obtained similar results when he asked graduating college seniors to predict the success of their upcoming job searches. These demonstrations are consistent with the idea that the neglect of alternative hypotheses might be a source of overconfidence (also see Griffin, Dunning, & Ross, 1990).

However, one problem is that the reason-listing manipulation used in previous experiments is somewhat heavy-handed. Asking subjects in a psychology experiment to list reasons why

Derek J. Koehler, Department of Psychology, Stanford University.

Preparation of this article was supported by a National Defense Science and Engineering Graduate Fellowship. I am indebted to Amos Tversky and Daniela O'Neill for their advice throughout the course of this research, and to Lyle Brenner for his help in analyzing the results of Experiment 4. I also thank Charles Gettys, Asher Koriat, and two anonymous reviewers for their helpful comments. The cooperation of the Psychology Department at San Jose State University in running Experiment 3 is gratefully acknowledged.

The experiments reported in this article served as partial fulfillment of the requirements for the PhD degree at Stanford University. I thank my dissertation committee, Ron Howard (chair), Lee Ross, David Rumelhart, Claude Steele, and Amos Tversky, for their time and their many suggestions.

Correspondence concerning this article should be addressed to Derek J. Koehler, who is now at the Medical Research Council Applied Psychology Unit, 15 Chaucer Road, Cambridge, CB2 2EF, England.

they might be wrong could act as a strong suggestion to give lower probability estimates, even if the subjects do not actually feel any less confident. From a slightly different view, the request that subjects produce reasons for some alternative might be taken as information implying that there actually are good reasons supporting the alternative, which have been overlooked. One of the purposes of the present experiments was to provide further evidence that overconfidence is caused in part by the neglect of alternatives through the use of a methodology that is not open to this argument.

The other impetus for these experiments was a desire to compare what happens when people are asked to generate their own hypotheses with what happens when they are merely asked to evaluate a hypothesis that is provided to them. Typically, in studies of confidence, subjects have been asked to select among a set of prespecified alternatives (often just two) in a multiple-choice format before making their confidence estimates. In day-to-day judgment, however, people often must produce their own initial set of possibilities before deciding which one is most likely to be correct.[1]

Gettys and his colleagues (e.g., Gettys & Fisher, 1979; Gettys, Mehle, & Fisher, 1986) are the only researchers to have examined in any detail the relationship between confidence and accuracy in tasks requiring subjects to generate their own hypotheses. They have produced considerable evidence of the fallibility of the hypothesis generation process, both in terms of people's ability to generate a complete set of possibilities and in terms of their ability to accurately assess the probability of the hypotheses they do generate. Specifically, these researchers have found that people tend to overestimate the probability of the hypotheses they generate (Mehle, Gettys, Manning, Baca, & Fisher, 1981) and, correspondingly, tend to underestimate the probability that the correct hypothesis is included in the remaining set of unspecified hypotheses (Gettys et al., 1986). As a result, the process of hypothesis generation itself is usually stopped too soon (e.g., Gettys & Fisher, 1979). Gettys and his colleagues have accounted for the overconfidence they observed in the hypothesis generation task using an availability-based model in which the inability to produce a complete set of possibilities is directly translated into underestimation of those possibilities that remain unspecified (cf. Fischhoff, Slovic, & Lichtenstein, 1978).

With one exception (considered in the General Discussion section), Gettys and his colleagues have focused on the hypothesis generation task and have not compared hypothesis generation with hypothesis evaluation. There are reasons to believe that having to come up with one's own hypotheses might lead to systematic differences in the way subsequent confidence estimates are made. For example, previous research (e.g., Anderson, Lepper, & Ross, 1980; Ross, Lepper, Strack, & Steinmetz, 1977; Sherman, Zehner, Johnson, & Hirt, 1983) has shown that asking subjects to produce an argument or explanation supporting some hypothesis leads them to express greater confidence that the hypothesis is true. One obvious speculation is that asking people to generate a hypothesis will have a similar effect.

On closer scrutiny, however, it seems that the two tasks require quite different approaches. Providing a supporting argument or explanation increases confidence because it draws attention to the focal hypothesis and requires the individual to temporarily treat it as if it were true (Koehler, 1991). Generating a hypothesis, in contrast, requires the individual to consider multiple hypotheses in the process of producing one that seems most likely to be correct. Consequently, people might express less confidence in a hypothesis they generate than in one that is presented to them for evaluation. This prediction is based on the notion of neglecting alternative hypotheses: If people who are asked to generate a hypothesis consider several before settling on one they favor (cf. Fisher, Gettys, Manning, Mehle, & Baca, 1983), they should consider more alternative hypotheses than people who are merely asked to evaluate a hypothesis that is presented to them. When evaluating someone else's hypothesis, in contrast, these alternatives are less likely to be considered and, as a result, the evaluator may be more confident than the person who generated the hypothesis in the first place.

To test this idea, I asked subjects in the first two experiments to predict the winners at the upcoming Academy Awards ceremony. Experiment 1 was conducted before the Academy of Motion Picture Sciences released its list of nominees. Experiment 2 was conducted after the five nominees in each category had been announced. The hypothesis was that subjects who generated their own predictions would express less confidence in them than others who evaluated these same predictions, but only in Experiment 1, before all the alternatives (nominees) were known.

Experiment 1

Method

Subjects. The 36 subjects in Experiment 1 participated in exchange for course credit in their introductory psychology class at Stanford University. Data from an additional 11 subjects were discarded because these subjects failed to provide the requested three nominees per category.

Procedure. Subjects were asked to complete a questionnaire in which they were to predict the Academy Awards winners. These predictions were made approximately 3 weeks before the list of nominees was released by the Academy. The first 12 subjects who participated in the experiment were assigned to the generate condition. They were asked to produce a list of three possible winners for each of three Academy Awards categories: Best Film, Best Actress, and Best Actor. All subjects considered the three categories in this order.

After producing each list of three candidates, subjects were asked to estimate the probability that their list contained the film, actress, or actor that would actually win. The instructions informed subjects that their probability estimates could range from 0% to 100%, where "0% represents no chance that the actual winner is included in the list and 100% represents certainty that the actual winner is included." Subjects were instructed to make their probability estimates so that over a great number of such estimates, approximately 40% of the lists assigned a 40% estimate should include the correct outcome, and so forth.

Two subjects in the evaluate condition were paired randomly with each subject in the generate condition and were asked to evaluate the

[1] Obviously, it is not the case that people first generate all possibilities and only then evaluate them; instead, presumably, people evaluate and monitor possibilities as they come to mind. Logically, though, any given hypothesis must be generated before it can be evaluated.

lists of possible winners the previous subject had produced. The 24 subjects in this second condition were told that a previous group had been asked to list three candidates they thought were likely to win in each category and that the responses of a subject chosen at random from this group were presented on their questionnaire. The responses of the previous subject were copied, by hand, onto the questionnaires completed by subjects in the evaluate condition. Subjects in the evaluate condition were asked to estimate the probability that each list contained the film, actress, or actor that was going to win. The instructions regarding the probability estimates were identical to those provided to the subjects in the generate condition.

Results and Discussion

Mean probability estimates for subjects in each condition are presented by Academy Awards category in the left half of Table 1. Overall, people expressed greater confidence in predictions made by someone else ($M = 66\%$, $SD = 18$) than they did in their own predictions ($M = 47\%$, $SD = 16$), $F(3, 32) = 4.55, p < .01$, by Hotelling's T^2 test. Breaking down the results by category indicates that subjects in the evaluate condition gave significantly greater probability estimates than did subjects in the generate condition when making predictions about the Best Actress and Best Actor categories, $t(34) = 2.58$ and 3.41, respectively, both $ps < .02$. The difference between probability estimates for the Best Film category, although in the predicted direction, was not statistically significant, $t(34) = 1.35, p > .15$.

Mean accuracy in the experiment was 33%. Compared with this level of accuracy, both groups in this experiment were overconfident, but the predictions of the subjects in the generate condition were less overconfident than those in the evaluate condition. The fact that only three events were considered limits the usefulness of this analysis, however. Experiment 4 allows the question of accuracy to be addressed more thoroughly.

The results of Experiment 1 are inconsistent with any model of judgment in which probability is solely a function of the event or hypothesis under consideration. If each person had a fixed probability estimate for any given hypothesis, then to the extent that people disagree in their judgments, randomly assigning hypotheses for evaluation should lead to lower probability judgments than should letting people select their own preferred hypothesis. Instead, subjects who produced the lists of candidates gave lower probability judgments.

I propose the following interpretation of these results. When asked to generate a hypothesis, people consider a greater number of alternative hypotheses than they do when asked to evaluate a prespecified hypothesis. When assessing the probability of the focal hypothesis, the additional alternatives considered in the hypothesis generation task are still salient and lead the judge to make more conservative probability estimates in allowance for such alternatives. People in the evaluate condition may have looked at the proposed candidates and based their probability estimates on how closely the candidates resembled the prototypical Oscar winner (i.e., they based their judgments on representativeness; see Kahneman & Tversky, 1972, 1973). Because most of them did resemble Oscar winners (e.g., Meryl Streep, Robert DeNiro, and the film *Awakenings* all seem like Oscar winners, although none of them in fact won), people in this condition gave relatively high probability estimates. The evaluation task apparently draws attention to the focal hypothesis and away from possible alternatives.

Table 1
Mean Confidence in Oscar Winner Predictions in Experiment 1 (With Unspecified Alternatives) and Experiment 2 (With Specified Alternatives)

	Condition and number of subjects			
	Experiment 1: Unspecified alternatives		Experiment 2: Specified alternatives	
Category	Generate ($n = 12$)	Evaluate ($n = 24$)	Choose ($n = 19$)	Precircled ($n = 19$)
Best film	64%	74%	86%	75%
Best actress	28%	50%	71%	63%
Best actor	49%	75%	77%	59%
M	47%	66%	78%	66%

If a difference in the number of alternative hypotheses considered is the correct interpretation of these results, then we would expect the discrepancy between assessments of one's own and someone else's hypothesis to disappear when all the possibilities are presented to subjects before they make their confidence estimates. Because both groups would have the complete set of possibilities in front of them when making their judgments, they would no longer differ in terms of the number of possible alternatives considered. Alternatively, it could be that subjects believe "two heads are better than one" and simply feel more confident in a judgment that is (in some sense) the joint product of two judges. From a slightly different view, it could be that subjects doubt their ability to perform the task and believe that they know less than the average subject. Finally, the fact that a previous subject had selected a certain hypothesis might have been treated as information suggesting that the hypothesis is true. These interpretations imply that the difference between evaluations of one's own and another person's hypothesis should hold even when all the alternatives are known. With this in mind, a second experiment was conducted after the Academy announced the Oscar nominees.

Experiment 2

Method

Subjects. A total of 38 subjects completed a short questionnaire while eating lunch at Stanford University's student union. The responses of 2 additional subjects were left unpaired with responses from the corresponding experimental condition and are not considered in the analyses that follow. Adding their data does not significantly affect the results.

Procedure. Experiment 2 took place approximately 2 weeks after the five nominees for each category had been announced by the Academy. All subjects were presented with the five nominees for each of the three categories (Best Film, Best Actress, Best Actor) used in Experiment 1.

Subjects in the choose condition were asked to circle the three (of the five) nominees that they believed were most likely to win in each category. (They were asked to circle three nominees rather than just

one nominee to make this second experiment as comparable as possible to the first.) The 19 subjects in this condition were told that their three circled nominees constituted their "choice set" for each category. After creating each choice set, subjects were asked to estimate the probability that it included the nominee that would actually win. They were told that their probability estimates should be "less than or equal to 100% (which represents absolute certainty that the choice set contains the winner)." Instructions regarding ideal calibration were similar to those used in Experiment 1 except that the example probability values used were higher.

There were 19 subjects in the precircled condition who were paired randomly with the 19 subjects from the choose condition and presented with the choices this previous group had made (along with the two other, uncircled nominees). Subjects in the precircled condition were asked to estimate the probability that each choice set presented to them included the nominee that would actually win. They were given the same instructions regarding probability estimation as were subjects in the choose condition.

Results and Discussion

In this second experiment, subjects who circled their own choices expressed greater confidence in them ($M = 78\%$, $SD = 10$) than did other subjects who evaluated these choices ($M = 66\%$, $SD = 11$), $F(3, 34) = 4.12, p < .02$, by Hotelling's T^2 test. As is shown in Table 1, this difference held both for the Best Film and Best Actor categories, $t(36) = 2.38, p < .03$ and $t(36) = 2.89, p < .01$, respectively. Although the difference between the two conditions was in the same direction for the Best Actress category, it was not statistically significant, $t(36) = 1.64, p > .10$. These results demonstrate that the difference between the generate and evaluate conditions in Experiment 1 are not attributable to a general tendency to express greater confidence in another person's hypothesis than in one's own.

If one were to compare results across the first two experiments (see Table 1), an interesting observation could be made. For subjects who were asked either to generate or else to circle their favorite candidates, prior knowledge of the five nominees had a large impact on confidence. Those subjects generating their predictions before the nominees were announced gave much lower probability estimates ($M = 47\%$) than did subjects who made their choices among the five nominees for each category ($M = 78\%$). This suggests that people in these conditions were fairly sensitive to the number of possible alternatives to the hypothesis under consideration.

In sharp contrast, subjects in the evaluate and precircled conditions appeared to be completely insensitive to the importance of knowing or not knowing the names of the five nominees. Subjects who were asked to evaluate the probability of some specified list of candidates were just as confident when they did not know the names of the nominees as when they did ($M = 66\%$), even though accuracy was bound to be greater in the latter condition. (In fact, accuracy increased from 33% to 54% once the nominees were known.) As would be expected if these subjects were basing their probability estimates on the representativeness of the candidate in question, the number of possible alternatives appeared to have no effect on their level of confidence.

Experiment 3

I conducted a third experiment to demonstrate a similar pattern of results using a new task, which involved guessing a person's occupation on the basis of a short character sketch. There were two additional reasons for this conceptual replication. First, it was important to ensure that the results of the first two experiments did not depend on the somewhat unusual procedure of asking people to generate or evaluate three alternatives rather than just one as is perhaps more typical. Second, it was necessary to demonstrate that the results were in no way affected by the 5-week time lag or the difference in experimental setting between Experiments 1 and 2.

Method

Subjects. A total of 122 students enrolled in an introductory psychology course at San Jose State University participated in the experiment in exchange for course credit. Of these, 6 subjects were not paired with subjects in a corresponding condition; their responses are not considered in the analyses that follow. There are only negligible differences when these data are added to the analyses.

Procedure. Subjects were presented with the following character sketch:

> Tim is 38 years old, industrious, and practical. He is married and has two children. The people he works with describe him as intelligent and always open to new ideas. His career has been quite successful. He is hard-working and careful; his friends often tease him for being a perfectionist in everything he does. This sometimes causes him to take things a bit too seriously, and as a result people who have just met him tend to think he lacks a sense of humor. His closer friends, however, recognize that he is actually a warm, cheerful man. Tim is fairly outgoing and has an active social life. For fun, he enjoys classic films and playing racquetball whenever he can find a game.

Subjects in the generate condition were asked to write their single best guess as to Tim's occupation. Subjects in the evaluate condition were presented with the guess of a randomly selected subject in the generate condition, yielding a total of 30 subject pairs. Subjects in the choose condition were asked to circle their preferred guess from a list of five possible alternative occupations. The five alternatives were chemical engineer, college professor, corporate lawyer, medical doctor, and stockbroker. Subjects in the precircled condition were paired randomly with subjects in the choose condition and evaluated the choice of the previous subject, yielding a total of 28 subject pairs.

All subjects were asked to estimate the probability that the guess under consideration was correct. The probability estimation instructions were essentially identical to those given in Experiment 1.

Results and Discussion

As predicted, subjects were more confident when evaluating someone else's hypothesis than when evaluating their own, but only when the alternatives were left unspecified. Subjects in the generate condition were significantly less confident in the accuracy of their guess ($M = 47\%$, $SD = 21$) than subjects in the evaluate condition ($M = 63\%$, $SD = 20$), paired $t(29) = 2.67, p < .02$. In contrast, subjects provided with specified alternatives were slightly more confident in the choose condition ($M = 51\%$, $SD = 20$) than were subjects in the precircled

condition ($M = 45\%$, $SD = 22$), paired $t(27) = 0.76$, *ns*.[2] Perhaps the most interesting result from this experiment is that the subjects presented with hypotheses for evaluation were significantly less confident when all the alternatives were specified in the precircled condition than when they were left unspecified in the evaluate condition, $t(56) = 3.09, p < .005$. Even though providing the entire set of possible alternatives could only improve accuracy, confidence was lower, presumably because the alternatives all seemed plausible and thus induced an increased sense of uncertainty.

Although Experiment 3 replicated the results of the two previous experiments in a single study, it does not address the question of accuracy. A fourth experiment was designed so that the relationship between accuracy and confidence could be examined in some detail. Are people who generate their own hypotheses more sensitive to their accuracy than people who simply evaluate these hypotheses, as suggested by my interpretation that they consider more alternatives, or is it the case that both groups are equally sensitive to the likelihood of the hypotheses, with the group that generates their own hypotheses simply giving constantly lower confidence ratings? To answer this question, a large number of judgments must be collected from every subject. To accomplish this, I turned, as have many previous researchers, to the domain of general knowledge questions.

Experiment 4

Method

Subjects. The subjects were 32 undergraduates enrolled in an introductory psychology course at Stanford University who participated in exchange for course credit.

Materials. A total of 120 general knowledge questions were selected randomly from the 300 questions included in Nelson and Narens's (1980) norms. These questions were originally constructed for use in feeling-of-knowing studies (e.g., Nelson, Leonesio, Landwehr, & Narens, 1986) and are well suited for the current task because they require a subject-provided answer rather than a choice among prespecified alternatives. The questions used in this sample vary greatly in difficulty, ranging from easy questions, such as "what is the name of the rubber object that is hit back and forth by hockey players?" (puck) to difficult questions, such as "what is the name of the town through which Lady Godiva supposedly made her famous ride?" (Coventry). Variation in item difficulty is desirable to measure subjects' sensitivity to the accuracy of their answers.

Procedure. Each subject was presented with all 120 questions. Each question was followed by a blank line in which the subject could write his or her answer and an 11-point probability scale running from 0% to 100%, in intervals of 10%.

The first 16 subjects in the experiment were assigned to the generate condition. These subjects simply wrote their best guess for each question and then estimated the probability that it was correct. Subjects were told, truthfully, that the purpose of the study was to examine whether people could accurately distinguish the questions they had answered correctly from those they had answered incorrectly. They were informed that each question had an answer that was exactly one word long, and that the questions varied greatly in difficulty. Their instructions regarding the confidence estimates were as follows:

> You should circle 0 only if you are completely certain that your answer is incorrect, and should circle 100 only if you are completely certain that your answer is correct. Circle 50 if you think that your answer is as likely as not to be correct. In general, you can think of your probability estimates in the following way. Of all the answers that you assign a 40 percent probability, about 40 percent of them should turn out to be correct. Similarly, if we looked at all the answers you assigned a 70 percent probability, it should turn out that about 70 percent of them should be correct. Thus, *not one* of the answers you assign a 0 percent probability should turn out to be correct, and *every single one* of your 100 percent probability assignments should be correct.

Subjects were informed that spelling errors were acceptable and that they need not adjust their probability estimates to compensate for the possibility of a spelling mistake. Subjects were encouraged to answer every question, even if they had to guess. This was done so that they would not adopt the strategy of only answering the questions they were certain they could answer correctly, which would effectively eliminate the variance of the probability estimates.

The remaining 16 subjects, assigned to the evaluate condition, were paired randomly with subjects in the generate condition. They were told that a previous group of subjects had answered 120 general knowledge questions and that in their packet they would find the answers provided by one of these subjects, with whom they had been randomly paired. They were also told that the study was designed to examine whether people could accurately distinguish correct from incorrect answers. The instructions given to the subjects in the generate condition were described to them, including the important fact that the initial subjects had been strongly encouraged to make their best guess even when they were not certain of the answer and the fact that correct spelling was unimportant. They were given essentially the same instructions regarding the probability estimation task as were the original subjects.

Results and Discussion

Although they were encouraged to guess when they were uncertain of an answer, some subjects in the generate condition did not provide answers to a considerable number of items, presumably because of their difficulty. Each subject answered an average of 96 (or 80%) of the 120 general knowledge questions. The items with missing responses were crossed out on the forms provided to the subjects in the evaluate condition, who were instructed to simply ignore them and move on to the next item. The results that follow include only those items that were completed.

I computed accuracy scores for each subject by noting the proportion of correct responses among the items that he or she answered. Mean subject accuracy was 55% ($SD = 12$); the highest accuracy achieved by a subject was 84% and the lowest was 38%. As expected, the items varied substantially in difficulty ($SD = 31$): eight items were correctly answered by all the subjects and 11 were not correctly answered by any of the

[2] At first glance, the results in the choose and precircled conditions may appear to be in conflict with earlier work (Ronis & Yates, 1987; Sniezek, Paese, & Switzer, 1990) in which subjects were found to be more confident when they evaluated two-alternative general knowledge questions with one alternative precircled than when they first circled their best guess before giving a confidence estimate. However, the common finding in these previous studies was that confidence in whichever alternative was preferred (as inferred by whether the confidence estimate was greater or less than 50% in the precircled condition) was greater when one alternative was precircled. In the current studies, the results are analyzed in terms of the circled rather than the preferred alternative.

subjects. Accuracy for items falling between these two extremes was uniformly distributed.

Mean confidence scores were also computed for each subject. As in the earlier studies, mean confidence in the evaluate condition ($M = 68\%$, $SD = 12$) was significantly greater than mean confidence in the generate condition ($M = 59\%$, $SD = 12$), $t(15) = 3.44, p < .005$. In this general knowledge task, as well as in the previous judgment tasks, people presented with hypotheses to be evaluated expressed greater confidence in their accuracy than did the people who produced these hypotheses in the first place.

The main purpose of this study was to determine whether subjects who generated their own answers would give more accurate confidence estimates than would other subjects who evaluated these answers. First consider the difference between mean confidence and accuracy. Subjects in the evaluate condition were clearly overconfident relative to their mean accuracy (mean difference = 13%). This degree of overconfidence is similar to that found in other studies posing general knowledge questions to subjects without performance feedback (e.g., Lichtenstein & Fischhoff, 1980). Subjects in the generate condition, in contrast, overestimated their accuracy by less than 4%, a finding that is rare in these studies.

Although the aggregate results are informative, it is desirable to compute some measure of the accuracy of each subject's set of confidence estimates. The standard measure of error used in studies of confidence is the Brier score, a quadratic error measure that can be decomposed into several interpretable components (for discussion, see Yaniv, Yates, & Smith, 1991). The Brier score ranges from 0 to 1, with lower scores indicating better performance. Brier scores were computed and decomposed separately for each subject in the experiment. As measured by the Brier score, confidence assessments were significantly more accurate in the generate condition ($M = .103$) than in the evaluate condition ($M = .176$), $t(15) = 3.79, p < .001$. Given this finding, I now turn to the decomposition of the Brier score to investigate the way in which the generate condition judgments are superior.

Calibration is a measure of the discrepancy between the probability assigned to a group of judgments and the mean accuracy of those judgments. Lower calibration scores indicate better performance. Perfect calibration is achieved if the mean accuracy for the group of 40% judgments is 40%, and so on. The calibration component of the Brier score indicates that subjects in the generate condition ($M = .106$) are better calibrated than subjects in the evaluate condition ($M = .300$), $t(15) = 2.60, p < .02$.

Resolution is a measure of how well the judge can separate correct answers from incorrect answers, and it is highest when judgments assigned to different confidence categories are maximally different in their accuracy. Perfect resolution is achieved only when the judge assigns all correct answers to one category and all incorrect answers to another. Resolution is an ideal measure with which to examine the current hypothesis that subjects in the generate condition are more sensitive to the differential accuracy of their answers than are subjects in the evaluate condition. Resolution is indeed greater for subjects in the generate group ($M = .157$) than for subjects in the evaluate group ($M = .104$), $t(15) = 3.36, p < .005$. Subjects in the generate group did not merely give lower probability judgments; their judgments were more sensitive to the accuracy of their answers.

Liberman and Tversky (1993) offered an ordinal performance measure that does not require that people's confidence estimates equal their accuracy. It measures whether judgments assigned to categories indicating greater confidence tend to be more accurate than judgments assigned to lower categories (regardless of the numerical probability estimate associated with the category). That is, it measures the degree to which the binary outcome variable can be described as a (weakly) monotonic function of the judgment. This measure is a variant of Goodman and Kruskal's (1954, 1959) gamma with an adjustment for ties that is well-suited to the confidence assessment task. Higher scores indicate better performance. A perfect score of 1 is achieved if there exists a value such that all judgments assigned a greater probability are correct and those assigned a lower probability are incorrect. Use of this monotonicity measure, again computed separately for each subject, also indicates more accurate confidence estimates for the generate subjects ($M = .851$) than for the evaluate subjects ($M = .685$), $t(15) = 3.06, p < .01$.

It is clear from these results that generating one's own hypothesis leads to more appropriate confidence assessments than does evaluating a prespecified hypothesis. Assuming that, as I have proposed, the hypothesis generation process brings to mind alternative hypotheses not considered in the evaluation task, these results are consistent with previous arguments that poor correspondence between confidence and accuracy arises from the tendency to neglect alternatives to the focal hypothesis. The present results suggest, among other things, that asking a judge to generate a hypothesis before evaluating its probability might be used as a "debiasing" technique.

Experiment 5

I have interpreted the results of the previous experiments as supporting the idea that when someone produces a hypothesis, more alternative hypotheses are considered and confidence is reduced as a result. Although contrasting the generation task with the evaluation task has provided supporting evidence, it is desirable to use some other experimental manipulation that avoids any possible alternative interpretation hinging on the difference between examining one's own hypothesis and examining someone else's. In Experiment 5, I asked subjects to consider their own hypotheses either immediately after generating them or after completing a distractor task. If the proposed interpretation is correct, subjects who evaluate their own hypotheses after a filled delay—like subjects asked to evaluate another person's hypothesis—should express greater confidence than they do immediately after generation because the delay should reduce the salience of alternatives brought to mind during the process of producing the hypothesis.

Method

Subjects. Subjects were 73 students enrolled in an introductory psychology course at Stanford University who participated in exchange for course credit.

Procedure. All subjects were given a questionnaire containing four personality descriptions (including that of Tim, described in Experiment 3). For each profile, subjects in the generate condition ($n = 24$) made their best guess as to the occupation of the person described and then estimated the probability that their answer was correct. Subjects in the evaluate condition ($n = 25$) were presented with the guesses made by a randomly chosen subject from the generate condition and evaluated the probability that each guess was correct. Subjects in the generate-delay condition ($n = 24$) initially made their occupation guesses without giving probability estimates and then, after a filler task that took approximately 15 min, estimated the probability that each guess was correct. The filler task was a questionnaire in which subjects were asked to select among a number of applicants in a hypothetical university admissions task.

Results and Discussion

In both the generate and evaluate conditions, half of the subjects completed the questionnaire before the filler task and half completed it after the filler task. This variable had no significant effect in either condition and so will not be considered further.

As predicted, mean confidence in the generate-delay condition ($M = 47.5\%$, $SD = 20$) was essentially identical to that in the evaluate condition ($M = 48.0\%$, $SD = 16$), multivariate contrast $F(4, 67) = 1.60$, *ns,* and was significantly greater than that in the generate condition ($M = 34.8\%$, $SD = 15$), multivariate contrast $F(4, 67) = 2.71, p < .04$, omnibus multivariate analysis of variance (MANOVA) $F(8, 136) = 2.14, p < .04$. Univariate analyses of each target profile showed that the general effect was statistically significant for three of the profiles and followed the same pattern for the fourth profile. In this experiment, then, the lowered confidence found in the generate condition was shown to disappear if a delay is inserted between the time of hypothesis generation and hypothesis evaluation. In other words, subjects became more confident if they evaluated their hypothesis after a delay rather than immediately after generation, supporting the present interpretation that alternatives considered during hypothesis generation cause lowered confidence in the focal hypothesis immediately after it is generated. Following a delay, however, the alternatives brought to mind by the generation task are less salient, and confidence increases as a result.

Experiment 6

Experiment 6 was a replication of Experiment 5, with an added evaluate-delay condition. This new condition served to test an alternative interpretation of the results of Experiment 5, claiming that a delay before probability estimation might generally increase confidence. Experiment 6 was designed to show that the delay increases confidence only after hypothesis generation.

Method

Subjects. Subjects were 118 students enrolled in an introductory psychology course at Stanford University who participated in exchange for course credit.

Procedure. For subjects in the generate condition ($n = 30$), the evaluate condition ($n = 30$), and the generate-delay condition ($n = 29$), the procedure was identical to that of Experiment 5, except that a different filler task was used. The filler task, which again took approximately 15 min to complete, was a series of 40 analogy problems taken from Sternberg's (1989) study guide to the Miller Analogies Test, modified to be "fill in the blank" rather than "multiple choice" problems. Subjects in the evaluate-delay condition ($n = 29$) were presented with the guesses made by a randomly chosen subject from the generate-delay condition; then, after the distractor task, the subjects evaluated the probability that each guess was correct.

Results and Discussion

The results closely matched those of the previous experiment. Mean confidence (across target profiles) was significantly lower in the generate condition ($M = 34.5\%$, $SD = 17$) than in the evaluate condition ($M = 45.1\%$, $SD = 24$), repeated measures $F(4, 26) = 2.70, p = .05$. As in Experiment 5, mean confidence in the generate-delay condition ($M = 46.5\%$, $SD = 25$) was essentially identical to that in the evaluate condition and was greater than that in the generate condition, $F(4, 54) = 2.38$, $p = .06$, by Hotelling's T^2 test. Mean confidence in the evaluate-delay condition ($M = 44.4\%$, $SD = 25$) was essentially the same as that in the evaluate condition, indicating that the delay itself (in the absence of the generation task) does not cause a general increase in confidence. This rules out the alternative interpretation of Experiment 5.

General Discussion

The main finding in the present study is that subjects who generate their own hypotheses express less confidence in their truth than do subjects who are presented with the same hypotheses for evaluation. This result appears to be robust, having been demonstrated in tasks that vary from real-world prediction (Experiments 1 and 2) and social inference (Experiments 3, 5, and 6) to general knowledge questions (Experiment 4). Because people generally give overconfident assessments in such tasks, it is not particularly surprising, given the main finding, that the mean confidence of subjects generating their own hypotheses is generally closer to their mean accuracy than is that of subjects asked to evaluate the hypotheses. However, as shown in Experiment 4, closer examination shows that people who have generated their own hypotheses make confidence assessments that are actually better calibrated and that more accurately discriminate correct from incorrect judgments. People appear to be better able to judge the accuracy of a hypothesis if it is one that they have just generated than if it is one presented to them for evaluation.

As mentioned in the introduction of this article, a study by Mehle et al. (1981) compared the effects of providing subjects with hypotheses with the effects of asking subjects to generate their own hypotheses instead. The main goal was to determine whether overconfidence in specified hypotheses could be reduced by drawing subjects' attention to alternative hypotheses contained in the unspecified "catchall" or residual hypothesis. The first experiment showed that both providing example hypotheses from the catchall and asking subjects to generate their own examples reduced the overconfidence in specified hypotheses observed in a control condition. The probability

assigned to the catchall set was somewhat greater when the example hypotheses were self-generated than when they were provided by the experimenter. In a second experiment, subjects were asked to directly estimate the probability of the generated or provided hypotheses. There was no difference in judged probability between the two conditions. Although these results may seem to contradict those reported here, in that the provided hypotheses did not receive higher probabilities than the generated hypotheses, it should be noted that, unlike the current study, different hypotheses were evaluated under the two conditions in the Mehle et al. (1981) study. Had the subject-generated hypotheses been presented to another group for evaluation, the current findings imply that they would have received higher probability estimates. The fact that the provided hypotheses were assigned equal or lower probabilities than the subject-generated hypotheses suggests that the provided hypotheses did not seem plausible to subjects (even though, in fact, they had higher veridical probabilities than the self-generated hypotheses).

I have argued that confidence is lower in the generation task than in the evaluation task because the process of producing the focal hypothesis brings more alternative hypotheses to mind. Two experimental manipulations provide evidence that is consistent with this interpretation and rule out the possibility that the difference between the generation and evaluation tasks arises from a general tendency to express greater confidence in someone else's hypothesis than in one's own. First, the difference between assessments of one's own and someone else's hypothesis disappears when a closed set of alternatives is provided. Second, the difference is also eliminated when a filled delay is inserted between the time of hypothesis generation and the time of probability estimation. The current study, then, supports the idea that people fail to consider alternatives they are capable of generating when evaluating a hypothesis that is presented to them, and the study does so in a way that avoids some of the problems associated with asking subjects why they might be wrong. In this final section, I briefly discuss some implications of these findings for the study of probability judgment.

The major finding here is that the same hypothesis can receive systematically different probability assignments depending on whether the hypothesis was generated or provided for evaluation. Normatively, of course, the probability value is supposed to reflect the likelihood of the focal event relative to all other possible events and thus should not change as a function of how many alternatives are explicitly considered. In practice, the requirement that the probability estimate be normalized relative to all possible alternatives—including those that do not readily come to mind—is psychologically unfeasible in many cases. The work of Gettys and his colleagues (e.g., Gettys et al., 1986) illustrates the difficulty of meeting this requirement in a complex task.

The proposal that judged probability depends on the number of alternatives to the focal hypothesis considered by the judge suggests that the judge's representation of an event plays a significant role in determining its perceived probability. Tversky and Koehler (1993) have recently developed a theory in which probability judgments depend on the description of the focal and the alternative hypotheses. The major assumption is that an explicit description that lists the component events included in the hypothesis receives greater psychological weight than does an implicit description of the same hypothesis. Thus, to the extent that the alternative hypothesis is described explicitly, the focal hypothesis is judged as less probable (cf. Fischhoff et al., 1978).

From this view, the results of the current study can be attributed to a difference in the way the alternative hypothesis is represented in the generation and evaluation tasks. Subjects in the evaluate condition are assumed to represent the alternative to the focal hypothesis as a more or less undifferentiated entity (i.e., an "all others" category representing the negation of the focal hypothesis). Subjects in the generate condition, on the other hand, are assumed to represent the alternative as a disjunction of the specific hypotheses that came to mind in the generation task and an all others category. The decomposition of the alternative hypothesis in this condition gives it greater weight and thus reduces confidence in the focal hypothesis. Although the current study provides a good start, further research is obviously needed to establish that it is in fact the consideration of additional hypotheses and not some other consequence of the generation task that leads to decreased confidence.

The proposed interpretation of the current results suggests that because the generation task induces elaboration of the alternative hypothesis, more accurate assessments of confidence may be elicited from a judge who is first asked to generate a set of hypotheses for consideration. Presenting specified hypotheses for evaluation can yield confidence assessments that are inferior to those the judge is capable of making and thus can fail to capitalize fully on the knowledge the judge has to offer.

References

Anderson, C. A., Lepper, M. R., & Ross, L. (1980). Perseverance of social theories: The role of explanation in the persistence of discredited information. *Journal of Personality and Social Psychology, 39,* 1037–1049.

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1978). Fault trees: Sensitivity of estimated failure probabilities to problem representation. *Journal of Experimental Psychology: Human Perception and Performance, 4,* 330–344.

Fisher, S. D., Gettys, C. F., Manning, C., Mehle, T., & Baca, S. (1983). Consistency checking in hypothesis generation. *Organizational Behavior and Human Performance, 31,* 233–254.

Gettys, C. F., & Fisher, S. D. (1979). Hypothesis plausibility and hypothesis generation. *Organizational Behavior and Human Performance, 24,* 93–110.

Gettys, C. F., Mehle, T., & Fisher, S. (1986). Plausibility assessments in hypothesis generation. *Organizational Behavior and Human Decision Processes, 37,* 14–33.

Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association, 49,* 733–764.

Goodman, L. A., & Kruskal, W. H. (1959). Measures of association for cross classifications: II. Further discussion and references. *Journal of the American Statistical Association, 54,* 123–163.

Griffin, D. W., Dunning, D., & Ross, L. (1990). The role of construal processes in overconfident predictions about the self and others. *Journal of Personality and Social Psychology, 59,* 1128–1139.

Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology, 24,* 411–435.

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

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3,* 430–454.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80,* 237–251.

Koehler, D. J. (1991). Explanation, imagination, and confidence in judgment. *Psychological Bulletin, 110,* 499–519.

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

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? *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, England: Cambridge University Press.

Mehle, T., Gettys, C. F., Manning, C., Baca, S., & Fisher, S. (1981). The availability explanation of excessive plausibility assessments. *Acta Psychologica, 49,* 127–140.

Nelson, T. O., Leonesio, R. J., Landwehr, R. S., & Narens, L. (1986). A comparison of three predictors of an individual's memory performance: The individual's feeling of knowing versus the normative feeling of knowing versus base-rate item difficulty. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12,* 279–287.

Nelson, T. O., & Narens, L. (1980). Norms of 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.

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

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

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

Ross, L., Lepper, M. R., Strack, F., & Steinmetz, J. (1977). Social explanation and social expectation: Effects of real and hypothetical explanation on subjective likelihood. *Journal of Personality and Social Psychology, 35,* 817–829.

Sherman, S. J., Zehner, K. S., Johnson, J., & Hirt, E. R. (1983). Social explanation: The role of timing, set, and recall on subjective likelihood estimates. *Journal of Personality and Social Psychology, 44,* 1127–1143.

Sniezek, J. A., Paese, P. W., & Switzer, F. S., III. (1990). The effect of choosing on confidence in choice. *Organizational Behavior and Human Decision Processes, 46,* 264–282.

Sternberg, R. J. (1989). *How to prepare for the Miller Analogies Test* (5th ed.). Hauppauge, NY: Barron's Educational Series.

Tversky, A., & Koehler, D. J. (1993). *Support theory: A nonextensional representation of subjective probability.* Unpublished manuscript, Stanford University, Department of Psychology, Stanford, CA.

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

Received June 17, 1992
Revision received December 7, 1992
Accepted April 23, 1993 ■