

Overconfidence in Probability and Frequency Judgments: A Critical Examination

LYLE A. BRENNER AND DEREK J. KOEHLER

Stanford University

VARDA LIBERMAN

The Open University of Israel, Tel Aviv, Israel

AND

AMOS TVERSKY

Stanford University

The overconfidence observed in calibration studies has recently been questioned on both psychological and methodological grounds. In the first part of the article we discuss these issues and argue that overconfidence cannot be explained as a selection bias, and that it is not eliminated by random sampling of questions. In the second part of the article, we compare probability judgments for single events with judgments of relative frequency. Subjects received a target individual's personality profile and then predicted the target's responses to a series of binary questions. One group predicted the responses of an individual target, while a second group estimated the relative frequency of responses among all target subjects who shared a given personality profile. Judgments of confidence and estimates of relative frequency were practically indistinguishable; both exhibited substantial overconfidence and were highly correlated with independent judgments of representativeness. © 1996 Academic Press, Inc.

Studies of judgment under uncertainty have indicated that people are often overconfident. Overconfidence is manifested in various forms, such as nonregressive prediction (Kahneman & Tversky, 1973) and overly-narrow confidence intervals (Alpert & Raiffa, 1982), but much of the evidence for overconfidence comes from calibration studies. In a typical calibration experiment, subjects are presented with a series of general knowledge questions (e.g., Which river is longer,

the Amazon or the Nile?) or prediction problems (e.g., Who will win the election, the incumbent or the challenger?). For each question, subjects select one of the two answers and assess the probability that their answer is correct. A judge is said to be *calibrated* if his or her probability judgments match the corresponding relative frequency of occurrence. Specifically, among all answers to which the judge assigns a given probability (say, 75%), the judge is calibrated if 75% of these answers are in fact correct.

Studies of calibration have shown that people's confidence often exceeds their accuracy. (For reviews of the literature, see Keren, 1991; Lichtenstein, Fischhoff, & Phillips, 1982; McClelland & Bolger, 1994; Yates, 1990). Overconfidence is common, but not universal; it is typically eliminated and even reversed for easy questions (Lichtenstein & Fischhoff, 1977). This phenomenon, called the difficulty effect (or the hard-easy effect), is hardly surprising. Because overconfidence is *defined* as the difference between mean confidence and overall accuracy, we expect a negative correlation between overconfidence and accuracy on purely statistical grounds.

The overconfidence observed in calibration studies has recently been questioned on both psychological and methodological grounds. We first discuss these objections, and then investigate alternative interpretations of overconfidence by comparing assessments of unique events with estimates of relative frequencies.

METHODOLOGICAL ISSUES

Deliberate versus Random Selection of Items

Several authors, notably May (1986; see also Juslin, 1994), have pointed out that the selection of particularly

difficult or surprising items could lead to spurious overconfidence. Consider, for example, the question, "Which city is further north: Rome or New York?" Although there is nothing deceptive or misleading about this question, most people find the correct answer (Rome) quite surprising. Questions like this, in fact, may be highly diagnostic of knowledge of geography, but the inclusion of many such questions is bound to produce overconfidence. Conversely, the deliberate exclusion of difficult or surprising items is likely to produce underconfidence.

Following May's lead, Gigerenzer, Hoffrage, and Kleinbölting (1991) have proposed a theory of probabilistic mental models (PMM) that attributes overconfidence to a biased selection of questions: "If general knowledge questions were a representative sample from the knowledge domain, zero overconfidence will be expected.... However, general knowledge questions typically are not representative samples from some domain of knowledge, but are selected to be difficult or even misleading. . . . 'overconfidence bias' results as a consequence of selection, not of some deficient mental heuristics" (Gigerenzer, 1993, p. 304). The basic assumption of this theory is that the judge is properly calibrated for any representative sample of questions, easy or difficult, hence the only cause of overconfidence is a nonrepresentative selection of items. PMM makes two bold predictions. First, it predicts that random sampling of questions from a natural domain will eliminate overconfidence, regardless of item difficulty. Second, it predicts that random sampling will eliminate the hard-easy effect (Gigerenzer *et al.*, 1991, p. 512). Both predictions, however, turned out to be wrong.

Recent studies have shown that for questions of moderate or high difficulty (say, where accuracy is below 65%), overconfidence is consistently observed even when the questions are randomly sampled from some natural domain of knowledge. Griffin and Tversky (1992) selected at random pairs of states (e.g., Alabama, Colorado) and asked subjects to indicate which state is higher on some attribute (e.g., population), and to assess their confidence in each answer. These judgments yielded significant overconfidence for all attributes, contrary to the first prediction of PMM. For example, mean confidence for population questions was 75%, whereas the corresponding accuracy rate was 68%. Furthermore, overconfidence was more pronounced for the more difficult attributes, contrary to the second prediction of PMM.

Similar results have been reported by Juslin, Olsson, and Björkman (1995), who summarized some 25 data sets, all based on random sampling of items. Although average confidence across all studies was well calibrated, all five data sets in which accuracy did not ex-

ceed 65% exhibited overconfidence (mean = 7%), and all four data sets in which accuracy exceeds 78% exhibited underconfidence (mean = 5%), contrary to the first prediction of PMM. These data also provide strong evidence against the second prediction of PMM that for representative samples of items the correlation between confidence and accuracy will be zero. In contrast, the correlation, across data sets, between the degree of over/under confidence and the proportion of correct responses was .78. Further evidence that is inconsistent with PMM has been reported by Ariely, Zauberaman, and Wallsten (1995) who concluded that "subjects demonstrated overconfidence in both the perceptual and the knowledge domains under conditions in which (a) all statistical artifacts were ruled out, and (b) items (in the knowledge case) were randomly selected."

Evidently, the major (though not the sole) determinant of overconfidence is the difficulty of the questions, not the manner in which they are selected. Overconfidence increases with item difficulty, both for random and non-random selection of questions, and we know of no evidence showing that random sampling of questions reduces overconfidence over and above the effect of item difficulty. Although the deliberate selection of questions can produce spurious overconfidence, or anything else for that matter, random sampling of questions is not sufficient to eliminate overconfidence, contrary to PMM. Early assessments of the promise of this theory (Gigerenzer, 1993; McClelland & Bolger, 1994) should be revised in light of the failure of its two basic predictions.

Finally, the very notion of representative sampling of questions is highly problematic because it is unclear how to define the population of questions that constitutes a meaningful knowledge domain. As suggested by Keren and Van Bolhuis (1994), the replacement of almanac questions with more natural tasks such as predicting the stock market, the weather, or the outcomes of sporting events, may offer a more promising approach for investigating people's confidence in their knowledge. It is noteworthy that overconfidence has been observed in some natural prediction tasks in which item selection is not an issue. Examples include physicians' predictions of pneumonia (Christensen-Szalanski & Bushyhead, 1981), economists' quarterly forecasts of recession (Braun & Yaniv, 1992), amateurs' predictions of bridge contracts (Keren, 1987), or players' predictions of their opponents' moves (Griffin & Tversky, 1992).

Confidence as a Dependent versus an Independent Variable

Erev, Wallsten, and Budescu (1994; see also Dawes & Mulford, 1996) have noted correctly that plotting accuracy as a function of judged confidence (as is commonly

done in calibration studies) yields a different picture than that obtained by plotting judged confidence as a function of level of accuracy (whenever such data are available). Obviously, the two regression lines have different slopes. Consequently, these authors argue, the same data can exhibit both overconfidence and underconfidence depending on whether judged confidence is the dependent or the independent variable. We take issue with this interpretation. Although the two calibration curves may be quite different, the question of whether average confidence exceeds average accuracy is independent of whether accuracy is plotted as a function of confidence or vice-versa. Hence, the oldest and simplest measure of overconfidence—mean confidence minus overall accuracy—provides a valid criterion for diagnosing over/underconfidence that does not depend on the mode of analysis. If the same set of data appear overconfident in one analysis and underconfident in another, it is an indication that an inappropriate index of overconfidence (e.g., the slope of the regression line) has been used. The statistical models proposed by Erev *et al.* (1994; see also Brenner, 1995; Juslin *et al.*, 1995), based on true score plus error, may help account for the distributional characteristics of the data, but they are not necessary to diagnose the presence or absence of overconfidence.

Subjective Probability versus Relative Frequency

A persistent objection to the calibration procedure concerns the meaningfulness of the comparison of average confidence to the percentage of correct answers. As one author put it, "To compare the two means comparing apples and oranges" (Gigerenzer, 1991, p. 88). We do not wish to address here the psychological implications of the debate between the frequentist and the subjectivist interpretations of probability. We only wish to point out that the subjects in calibration experiments are explicitly instructed to use the scale such that, say, 70% of the statements to which they assign 70% confidence will, in fact, be valid. In several experiments subjects were actually paid according to this criterion. The lack of proper calibration, therefore, indicates that, at the very least, subjects fail to use the scale as instructed. This failure has significant implications because the probability scale is commonly used for communication. A client who is informed by her lawyer that her chances for winning the case are 99% may be justifiably upset to learn that when the lawyer expresses a 99% confidence he is actually correct only 75% of the time.

Aside from the normative question of whether average confidence should match the corresponding hit rate, there is an interesting empirical question: what happens when people estimate relative frequencies instead of single-event probabilities? To the best of our

knowledge, May (1987, 1988), was the first to report that while average confidence commonly exceeds the percentage of correct responses, people's estimates of the relative frequency of items that they have answered correctly are generally lower than the actual number. In her study of students' knowledge of psychology, the percentage of correct responses was 72%, average confidence was 81%, and the mean estimate of the percentage of correct responses was 63%. These data yield 9% overconfidence in judgment of probability and 9% underconfidence in judgment of relative frequency. This pattern has been replicated in subsequent studies (e.g., Gigerenzer *et al.*, 1991; Griffin & Tversky, 1992; Sniezek, Paese, & Switzer, 1990), although the degree of underconfidence varied considerably across studies.

These data indicate that people do not estimate the frequency of correct responses by averaging their confidence across all items. However, the data do not answer the question of whether the observed difference between the two tasks is due to the response scale (confidence versus frequency), or to the nature of the evidence on which the judgment is based. The following study explores this question.

To this end, we chose a task in which subjects can make either confidence judgments or frequency estimates on the basis of the same evidence. One group of subjects (the target group) completed a personality inventory in which they (a) rated themselves on three bipolar personality dimensions, and (b) answered a number of binary forced-choice questions regarding their behavior and preferences. A second group of subjects (the prediction group) was given the target's personality profile, and asked to predict the target's responses to the binary questions and to express their own confidence in these predictions. Half the subjects in this group predicted the response of an individual target, and the other half estimated the relative frequency of the response among all target subjects who share a given personality profile.

If, as suggested by Gigerenzer (1991), cognitive illusions disappear when subjects assess relative frequency instead of single-event confidence, we may expect overconfidence in the latter task but not in the former. On the other hand, if subjects' confidence in their predictions depends primarily on the degree to which the behaviors in question are representative of the target's personality profile (Kahneman & Tversky, 1973), we expect both groups to exhibit overconfidence.

EXPERIMENTAL DESIGN

The Target Group

The target group consisted of 206 undergraduates enrolled in an introductory psychology course at Stan-

ford University. Each completed an approximately 30-min personality inventory as part of a course requirement. Students were asked to rate themselves in terms of three global personality dimensions adapted from the popular Myers–Briggs Lifestyles Inventory. All target students rated themselves in terms of three bipolar dimensions defined as follows:

EXTROVERT—project energy outward; enjoy interaction with people, or

INTROVERT—keep energy inside; enjoy solitude.

ANALYTIC—prefer to act through a logical, step-by-step process, or

INTUITIVE—prefer to act on inspiration or imagination.

ADAPTIVE—seek to adapt my life to changing circumstances, or

DECISIVE—seek to control my life, exerting my will on events.

For each dimension, the subjects checked one of six boxes to indicate the extent to which one pole was a better description of their personality than was the other. We classified each subject in terms of one or the other pole based on whether the response fell above or below the mean self-rating on that dimension. In this manner, every target was classified as Extrovert or Introvert, as Analytic or Intuitive, and as Adaptive or Decisive, yielding eight possible personality *profiles*.

Subjects were then presented with 50 binary-choice questions involving dispositions, behavior, and preferences (e.g., Are you often late for class? Yes/No; Do you regularly make lists of things to be done? Yes/No; Do you enjoy gambling? Yes/No). The questions were selected to span a wide range of activities and attitudes. In addition, subjects were asked to choose between 28 pairs of potential occupations (e.g., accountant versus social worker; high school teacher versus high school principal). They were asked to indicate which of the two occupations better suited their personalities.

The Prediction Group

The prediction group consisted of 39 Stanford students recruited through an advertisement. They were paid \$7 dollars for their participation in the experiment, which lasted approximately 45 min. Prediction subjects were first acquainted with the three major personality dimensions used in the personality inventory. They were shown the same scales used by the Target Subjects, each pole of which was labeled with the appropriate term and description. They were told that a large group of students from their university had completed the personality inventory and that each had been classified as either Introvert or Extrovert, Ana-

lytic or Intuitive, and Adaptive or Decisive on the basis of whether their self-ratings on these dimensions were above or below the group mean. The subjects were told that their task would be to predict, on the basis of the personality profile, the responses of the target subjects to the rest of the personality inventory.

We selected two profiles for the prediction task: Extrovert, Intuitive, Decisive (EID) and Introvert, Analytic, Decisive (IAD). These were the two most common profiles that differed on two out of the three dimensions. Subjects predicted the responses for both personality profiles (IAD and EID); the order in which the two profiles were considered was counterbalanced. To prevent confusion or forgetting, the profile being considered appeared at the top of each page of the prediction questionnaire.

Each prediction subject served in one of two conditions. In the *individual* condition, subjects were asked to predict the responses of a single target subject. They were presented with a personality profile and were told that it described a single individual, specified by initials, who had been selected at random from all target subjects with that profile. Prediction subjects then predicted the target's responses for all the questions of the original inventory. Below each question was an 11-point scale (labeled "Probability") ranging from 50 to 100% in 5% intervals. Subjects circled a number indicating their confidence, or subjective probability, that their prediction was correct (i.e., that the target person actually chose the predicted option).

In the *aggregate* condition, respondents were asked to predict the percentage of target subjects of a given profile who chose a particular answer. For each question, they first predicted the response chosen by the majority of target subjects with a given profile (IAD or EID). They then circled a number on the 11-point scale (labeled "Percentage") ranging from 50 to 100% to indicate their best estimate of the percentage of these targets who chose the option in question.

For example, given the profile Extrovert Intuitive Decisive and the item "Do you enjoy gambling? Yes/No," prediction subjects in the aggregate condition were asked first to indicate whether the majority of EID targets responded "Yes" or "No" and then to estimate the size of this majority. Prediction subjects in the individual condition were asked first to indicate whether the particular randomly chosen EID target responded "Yes" or "No" and then to assess their confidence in this prediction.

Base Rate Estimates

We also gathered estimates of the response base rates for all questions. Subjects in the base rate group

were asked to estimate the overall percentage of their peer students that would choose a given alternative for each question, without any information regarding personality dimensions. Estimated base rates for each of the 50 questions were obtained from a group of 96 students enrolled in Introductory Psychology at Stanford University as part of a questionnaire packet handed out to the class. A different group of 39 Stanford students estimated base rate responses for the 28 occupational choices.

Representativeness Judgments

A separate group of 127 subjects were given the original personality inventory and were asked which of the two responses to each item was more representative of a given personality profile. Approximately half of the subjects evaluated the EID profile, and the remainder evaluated the IAD profile. For example, subjects were asked which profession, *accountant* or *social worker*, is more representative of the Extrovert, Intuitive, Decisive personality type. The percentage of subjects that selected each response is taken as a measure of the degree to which this response is representative of the respective profile. Thus, in addition to the targets' responses, we obtained three types of data from different groups of subjects: predictions, base rate estimates, and representativeness judgments.

RESULTS

Diagnosticity of the Personality Dimensions

To test whether the personality dimensions are predictive of the responses of the target group to the remainder of the inventory, we computed the correlations between each of the 78 items and each of the three binary personality dimensions. These values were significantly higher than expected by chance,¹ $\chi^2(78) = 220$, $p < .0001$. The correlations were generally consistent with lay expectations. For example, most Extroverts preferred to be with someone else when they are upset, while most Introverts preferred to be alone. People who described themselves as Analytic were more likely to make lists and less likely to have an artistic hobby than people who described themselves as Intuitive. Decisive subjects more often indicated strongly held political views and more often reported that they found it difficult to admit their mistakes than did Adaptive subjects.

¹ Each correlation was transformed using Fisher's z transformation, squared, then added together. Under the null hypothesis of no correlation, this sum follows a χ^2 distribution.

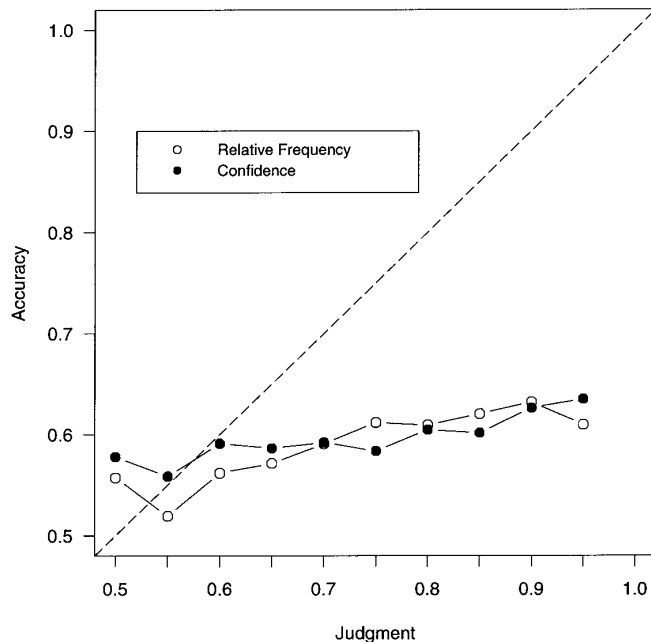


FIG. 1. Confidence and relative frequency calibration curves.

Individual versus Aggregate Prediction

Subjects in the individual condition (who assessed their confidence in the prediction of the response of a single individual) and subjects in the aggregate condition (who estimated the percentage of the predicted response in the target group) made essentially the same judgments. Comparisons between the two groups on the 78 items yielded 8 significant differences (at $\alpha = .10$) for the IAD profile and 10 significant differences for the EID profile, which is very close to what is expected under the null hypothesis of no difference.

Figure 1 presents the calibration curves for the two conditions. (Since there were only a few estimates of 100%, they are not included in the graph.) As noted above, there is practically no difference between judgments of confidence in individual predictions and estimates of relative frequency. Furthermore, both groups exhibited substantial overconfidence. Average confidence, across all items, was 72.1% whereas average accuracy was only 59.6%. The difference is highly significant, $t(36) = 6.5$, $p < .0001$. As expected, accuracy generally increases with judged confidence or relative frequency, but the relation is very weak. Because overconfidence is very much in evidence even in judgments of relative frequency, it cannot be attributed to the nature of the required response.

Correlational Analysis

Table 1 presents correlations (across all items) between the mean responses in the four tasks, separately

TABLE 1

Correlations between Prediction, Target, Base Rate, and Representativeness Means, across Judgment Items for the Two Personality Profiles

Correlation	Profile	
	IAD	EID
r (prediction, target)	.48	.55
r (base rate, target)	.61	.75
r (prediction, rep.)	.96	.89

for each profile. The results may be summarized as follows. First, there is a moderate correlation between the predictions and the target responses (.48 and .55), indicating that subjects were able to predict the targets' responses from the targets' personality profile with reasonable success. However, the fact that the targets' responses correlated even higher (.61 and .75) with the estimated base rates (provided by a different group) suggests that subjects did not make appropriate use of their own knowledge of this information. Thus, subjects could have achieved a higher predictive validity by ignoring the personality profile altogether and relying exclusively on their judgments of the relevant base rates. Obviously, combining both personality profile and base rate could produce even better predictions. Finally, subjects' predictions were highly correlated (.96 and .89) with the judgments of representativeness obtained from a separate group of subjects. Taken together, these data are consistent with the notion that subjects made their predictions on the basis of the degree to which the target's potential response is representative of the relevant personality profile, with insufficient regard for the base rate of the response.

We have translated the personality inventory into Hebrew and replicated the study at Tel Aviv University, with minor procedural variations. In the Israeli version, each prediction subject ($N = 86$) evaluated only one profile (EID or IAD) in the individual condition. The five most accurate subjects each received a \$20 prize. The results were essentially the same. Mean confidence was 76.9% whereas mean accuracy was 56.7%, $t(170) = 31.7$, $p < .0001$. The correlations between the predictions and the target responses were .63 and .62, for IAD and EID respectively; the correlation between the target response and the estimated base rate were .63 and .75; and the correlations between the predictions and representativeness were .93 and .90.

Inverse Regression

Note that, in standard calibration studies involving the prediction of unique events, accuracy is plotted as

a function of judged confidence because judgments must be aggregated into confidence categories before a measure of accuracy can be computed. However, in the aggregate condition of the present study we can also plot for each item the mean judgment of relative frequency against the actual relative frequency. Fig. 2 displays the scatterplot for all 78 items for the IAD profile (the plot for the EID profile is essentially the same). It also displays the regression lines for the prediction of actual frequency from judged frequency and vice versa. Obviously, the two regression lines have different slopes (as noted by Erev *et al.*, 1994), reflecting the low correlation between the observed and predicted proportions ($r = .22$). The two regression lines cross at the means of the two variables: mean estimate, 72%; mean relative frequency, 56%. This discrepancy reveals substantial overconfidence, independent of the regression lines. Overconfidence is also manifested by the observation that mean estimates exceed the actual proportions in 87% of the items. In the absence of overconfidence, an equal number of items are expected to fall above or below the identity line.

DISCUSSION

The results of the present study may be summarized as follows. First, subjects were able to predict the targets' responses from the targets' personality profiles with modest success. Second, subjects who made confidence judgments concerning unique events and sub-

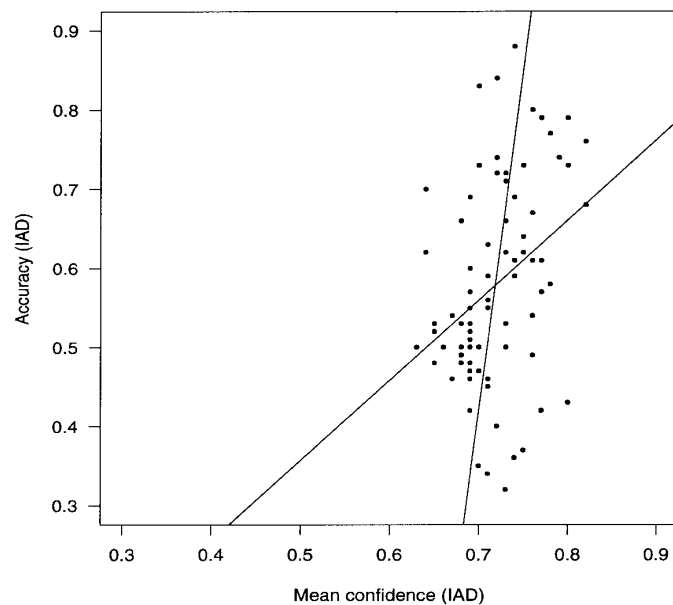


FIG. 2. Scatterplot of accuracy vs mean confidence, by judgment item.

jects who estimated relative frequencies produced essentially the same responses. Third, both confidence judgments and frequency estimates exhibited substantial overconfidence. Fourth, subjects did not make effective use of their knowledge of base rate information. Fifth, subjects' predictions were highly correlated with independent judgments of representativeness (see Koehler, Brenner, Liberman & Tversky, 1996).

The present data refute the claim that "eliminating the experimenter's normative confusion between single events and frequencies made the subjects' 'overconfidence bias' disappear" (Gigerenzer, 1991, p. 90). The replacement of confidence judgments by the estimation of relative frequencies did not eliminate overconfidence in this case. This result suggests that the pattern of overconfidence observed in judgments of single items and the underconfidence observed in the estimation of the percentage of correct answers (e.g., May 1987, 1988) is not due to the response scale (confidence versus frequency) but rather to the nature of the evidence on which the judgments are based. The present findings are consistent with Griffin and Tversky (1992):

A sportscaster, for example, can be asked to assess his confidence in the prediction of each game as well as the number of games he expects to predict correctly. According to the present account, these judgments are not expected to coincide because they are based on different evidence. A judgment of confidence in a particular case, we propose, depends primarily on the balance of arguments for and against a specific hypothesis, e.g., the relative strength of two opposing teams. Estimated frequency of correct prediction, on the other hand, is likely to be based on a general evaluation of the difficulty of the task, the knowledge of the judge, or past experience with similar problems. (p. 431)

Because in the present study both judgments of confidence and estimates of relative frequency were based on exactly the same evidence (i.e., a given personality profile), subjects made essentially the same judgments in the two tasks. This is generally not the case in studies of calibration in which the individual and the aggregate judgments are based on different evidence, as in the sportscaster example above (see also Kahneman & Lovallo, 1993).

In conclusion, it is evident that overconfidence is not eliminated by random selection of items, it does not disappear in estimates of relative frequency, and it cannot be treated merely as a regression artifact. Furthermore, comparable levels of overconfidence have been observed in studies of subjective probability distributions (see e.g., Alpert & Raiffa, 1982; Lichtenstein, Fischhoff, & Phillips, 1982), a task to which the above objections do not apply. On the other hand, the interpretation of the overconfidence observed in calibration experiments is complicated by the ubiquitous difficulty effect, which is as much a statistical as a psychological phenomenon.

REFERENCES

- Alpert, M., & Raiffa, H. (1982). A progress report on the training of probability assessors. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 294–305). New York: Cambridge University Press.
- Ariely, D., Zauberman, G., & Wallsten T. S. (1995, August). *Statistical and psychological considerations in calibration research: Overconfidence, the hard-easy effect, and the influence of response scale on judgment*. Paper presented at SPUDM-15, Jerusalem, Israel.
- Braun, P. A., & Yaniv, I. (1992). A case study of expert judgment: Economists' probabilities vs. base-rate model forecasts. *Journal of Behavioral Decision Making*, 5, 217–231.
- Brenner, L. A. (1995). *A stochastic model of the calibration of subjective probabilities*. Unpublished doctoral dissertation, Stanford University.
- 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, in 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.
- Gigerenzer, G. (1991). How to make cognitive illusions disappear: Beyond "heuristics and biases". In W. Stroebe & M. Hewstone (Eds.), *European Review of Social Psychology*, (Vol. 2, pp. 83–115). New York: Wiley.
- Gigerenzer, G. (1993). The bounded rationality of probabilistic mental models. In K. I. Manktelow & D. E. Over (Eds.), *Rationality: Psychological and philosophical perspectives*. London: Routledge.
- Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review*, 98, 506–528.
- Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology*, 24, 411–435.
- Juslin, P. (1994). The overconfidence phenomenon as a consequence of informal experimenter-guided selection of almanac items. *Organizational Behavior and Human Decision Processes*, 57, 226–246.
- Juslin, P., Olsson, H., & Björkman, M. (1995, August). *Format-dependence and multiple sources of bias in probability assessment: On the origin and nature of the stochastic components of judgment*. Paper presented at SPUDM-15, Jerusalem, Israel.
- Kahneman, D., & Lovallo, D. (1993). Timid choices and bold forecasts: A cognitive perspective on risk taking. *Management Science*, 39, 17–31.
- Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, 80, 237–251.
- 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.
- Keren, G., & Van Bolhuis, J. (1994). *On the ecological validity of calibration studies*. Unpublished manuscript, University of Technology, Eindhoven, The Netherlands.
- Koehler, D. J., Brenner, L. A., Liberman, V., & Tversky, A. (1996).

- Confidence and accuracy in trait inference: Judgment by similarity. *Acta Psychologica*, in press.
- 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., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 306–334). New York: Cambridge Univ. Press.
- May, R. S. (1986). Inferences, subjective probability and frequency of correct answers: A cognitive approach to the overconfidence phenomenon. In B. Brehmer, H. Jungermann, P. Lourens, & A. Sevooan (Eds.), *New directions in research on decision making* (pp. 175–189). Amsterdam: North Holland.
- May, R. S. (1987). *Calibration of subjective probabilities: A cognitive analysis of inference processes in overconfidence* (in German). Frankfurt: Peter Lang.
- May, R. S. (1988). Overconfidence in overconfidence. In A. Chikan, J. Kindler, & I. Kiss (Eds.), *Proceedings of the 4th FUR Conference*. Dordrecht: Kluwer.
- McClelland, A. G. R., & Bolger, F. (1994). The calibration of subjective probabilities: Theories and models 1980–1993. In G. Wright & P. Ayton (Eds.), *Subjective probability* (pp. 453–482). Chichester: Wiley.
- Sniezek, J. A., Paese, P. W., & Switzer, F. S. (1990). The effect of choosing on confidence in choice. *Organizational Behavior and Human Decision Processes*, 46, 264–282.
- Yates, J. F. (1990). *Judgment and decision making*. Englewood Cliffs, NJ: Prentice Hall.

Received: October 4, 1995