Does Overconfidence Lead to Poor Decisions? A Comparison of Decision Making and Judgment Under Uncertainty

Ed Bukszar, Simon Fraser University

Two within-subject studies of business executives indicate that overconfidence in judgment is reduced when actual decisions are made. Subjects projected quarterly earnings for 50 firms based on reported earnings from the previous 12 quarters. They stated their confidence in forecasts, were given a $10 allocation and were allowed to invest in any, all or none of their forecasts. Subjects chose to act on a relatively small portion of their forecasts but were more accurate and better calibrated when making investment decisions than when making forecasts. The relatively more accurate subjects made more investment decisions and riskier decisions than the relatively less accurate subjects. This suggests that subjects had a sense of whether their knowledge was appropriate for decision making, and acted accordingly. Improved calibration for decision making (investments) compared to judgments (forecasts) appears to have been the result of an additional evaluation stage which occurred between making a forecast and acting upon it. Most subjects tended to 'think twice' before acting, which may have lead to more thorough information processing and improved calibration.

Findings in the literature on overconfidence are wide-ranging and robust (Fischhoff, 1982; Lichtenstein, Fischhoff, & Phillips, 1982).[1] Poorly calibrated decision makers are thought to make ill-advised choices based on sub-optimal information searches. Generalizing somewhat, decision makers tend to be overconfident for difficult problems and underconfident for easy problems (Lichtenstein & Fischhoff, 1977). Findings are not unequivocal with respect to experts. Some show significant overconfidence while others show reasonably well calibrated

---

[1] The following can be thought of as a very brief primer regarding overconfidence, and the terminology related to its study.

When making an estimate of an event occurring, or the likelihood of being accurate in a forecast, or in an answer to any question, individuals can make an estimate of the likelihood that they are correct. For example, I am 90% certain that I am right when I say ... In this example, 90% is the estimate of **confidence** in making the correct statement or forecast. The higher the percentage, the higher the confidence.

It is possible to pool all such statements across all estimates. That is, all 90% estimates can be pooled and then compared to outcomes. An individual is well **calibrated** when his or her estimates match outcomes. In other words, 90% forecasts should be correct 90% of the time on average, 70% forecasts should be correct 70% of the time, etc. A perfectly calibrated individual would have estimates and outcomes coincide throughout the range of possible estimates, from 0 to 1.

Overconfidence occurs when estimates of accuracy are greater than actual accuracy. **Underconfidence** occurs when estimates of accuracy are less than actual accuracy.

experts. Characteristics of the decision tasks may account for the differences (Keren, 1991). Weather forecasters, accountants and loan officers can be trained to be calibrated due to the clarity and timeliness of feedback while physicians, psychologists and strategists are more prone to overconfidence due to the perceived uniqueness of the circumstances at hand and the difficulty of learning from feedback under these circumstances (Christensen-Szalanski & Busheyhead, 1981; Hogarth, 1981; Oskamp, 1965; Russo & Schoemaker, 1992; Tomassini, Solomon, Romney, & Krogstad, 1982).

Relatively unaddressed in this research is the question of whether there is a difference between overconfidence with respect to judgments and overconfidence in decisions. Most studies of overconfidence require subjects to evaluate information and state their confidence level in judgments they would make from the information. The findings of overconfidence in the reported judgments is taken as an indication that overconfidence would also be evident in decision making, that is, that overconfidence in judgment leads to ill-advised action. However, in many real-world circumstances, people are *not* required to act on their judgments. Action implies consequences for which decision makers are often held accountable. If they believe their knowledge is limited, individuals may restrict their decision making, that is, the items upon which they are prepared to act, to a subset of their judgments. (Clearly, in some instances, it is a decision to not act upon an item, particularly when non-action implies consequences. However, in general, I mean to restrict the category 'decision' to actions based upon judgments.)

For example, most of us have expressed an opinion rather strongly only to recant somewhat when asked to back it up with action, in the form of a bet or an investment. Timidity with respect to actual decisions is likely the result of loss aversion as decision makers have been shown to weigh losses at least twice as much as gains (Kahneman & Lovallo, 1993; Kahneman

---

The **Brier Score** is a measure of the deviation from perfect calibration, calculated in a manner not unlike the error in a regression function (Brier, 1950). Low Brier scores indicate good performance, whereas high scores indicate poor performance. The Brier score is computed as follows (from Yates, 1988).

Let fi denote the respondent's probability that her/his judgment i is correct, where i = 1, ... N, and N = the total number of judgments. For each of the i judgments, an outcome index di is defined such that

di = 1, if the judgment is correct,
= 0, if the judgment is incorrect.

The Brier score is the mean squared difference between the stated probability and the outcome index summed across all N judgments:

$$\text{Brier score} = \left(\frac{1}{N}\right)\sum_{i=1}^{N}\left(\mathrm{f}_i - d_i\right)^2 .$$

The lower the Brier score, the better the overall judgment performance. For a more complete discussion see Lichtenstein & Fischhoff (1977).

& Tversky, 1979). Faced with the possibility of a loss, decision makers 'think twice' before acting, resulting in somewhat cautious behavior. Overall, the result may be a "conjunction of biases" (Kahneman & Lovallo, 1993, p. 30) which produces offsetting errors, with timid decisions balancing, to varying degrees, overconfident judgments. In such cases, overconfidence in judgment does not translate to overconfidence in decision making. As well, individuals may take greater care when making decisions compared with judgments, leading to improved calibration (Paese & Feuer, 1991; Tetlock, 1983). The present study investigates these issues.

STUDY 1

Twenty-five executives, enrolled in a Strategic Management seminar in an EMBA program in Vancouver, British Columbia, participated in a decision making experiment. Subjects were from middle and upper management with mean age of 40. Ten of the 25 subjects (40%) were women.

Subjects were asked to forecast quarterly earnings per share for 50 firms, based on the previous 12 quarters of actual earnings. The firms were selected randomly from Valueline Investment Survey but were not identified by name; subjects were informed about this selection process to eliminate suspicion that firms had been chosen to make forecasting particularly difficult.

A forecast was deemed accurate if the actual earnings for the firm in the forecasted period fell within a range determined by taking the subject's forecast, plus or minus a value indicated for each firm. An example of what subjects actually saw when participating in this experiment follows:

Firm #	Yr	QI	QII	QIII	QIV	Earnings Est. 90-1	+/-
1	1987	25	28	30	27		
	1988	37	40	41	32		
	1989	40	43	44	36	_____	5

To walk the reader through the example, assume that a subject forecasted earnings of $45 per share in 90-1 for the firm above. Given the +/- figure of 5, if actual earnings fell within the range of $40-$50 (45-5 and 45+5), the subject's forecast would be deemed accurate.

The plus or minus value was determined by taking 15% of the mean of earnings for the 12 previous quarters; subjects were not informed as to the methodology for determining this value. To motivate all subjects to perform as well as possible, $50 and $20 cash rewards were offered for the two individuals with the greatest number of accurate forecasts.

In addition to making a forecast of earnings, subjects were asked to state their confidence in each forecast, from .1 to .9. Confidence assessments reflect the subjective probability that the actual earnings for the firm would fall within the range defined by the forecast and the plus/minus value for each firm. Confidence and calibration were explained to the subjects

both verbally and in writing. Subjects had no information regarding the confidence assessments of others.

After completing their forecasts and indicating their confidence in each, subjects were asked to list the forecasts that they wanted to invest in. By investing, subjects were in essence betting on their belief in the accuracy of the forecast. Thus, subjects were given a choice to either act or not act on each of the forecasts they made. To motivate subjects to act on their forecasts, subjects were informed at the beginning of the experiment that each would be given an account with $10 which they could use to invest in their forecasts. For example, by investing (betting) $5 on a forecast, a subject would either earn $5 if the forecast was accurate (that is, if actual earnings for the firm fell within the range as described above) or lose $5 if it was inaccurate (if it fell outside of the range). As such, investment payoffs worked very much like a common bet.

Subjects could invest in any increment but if they chose to invest, they had to invest all $10 of their funds. Subjects could make one, $10—investment or ten, $1—investments or combinations in between. Subjects could also make investment choices totaling more than $10. For example, a subject could list three, $5—investments. Thus subjects were not constrained by their initial allocation of funds; they could make as many investment choices as they saw fit. However, if the balance in their accounts hit zero, no other investment options were considered. So, in the example above, if the first two, $5—investments were inaccurate, no money would remain in the account, and the third investment would not be evaluated.

Subjects were advised that they could select as many investments as they deemed appropriate and that the investments would be evaluated as long as money remained in their accounts. Theoretically, a subject could bet on all 50 of his/her forecasts, and if accurate, could earn a substantial sum of money. Subjects were allowed to keep all of their earnings. They were instructed to make as much money as possible and were given two hours to complete the task. Subjects choosing not to invest were allowed to keep $5.

Upon completion of the experiment, investments were evaluated and accounts credited or debited as appropriate. Thus, no feedback was given regarding the accuracy of the forecasts during the course of the experiment. Comparisons between judgments (income forecasts) and decisions (the forecasts chosen for investment) are the focus of this study.

Results

Average confidence for all judgments was .44; forecast accuracy ranged from 20% to 38% with a mean of 29%. Thus, overall, subjects were overconfident in judgment.

The structure of the payoff rules encouraged investments in forecasts for risk-neutral individuals with confidence assessments as low as .5. There were 627 of these 'eligible' investment options out of the 1250 total (1250 = 25 subjects multiplied by 50 forecasts made by each subject). Subjects actually chose a total of only 165 investments, or 26% of the total eligible. To be more precise, subjects chose to invest in only 69% of the forecasts for which a .9 confidence level had been indicated. The proportions dropped to 27% for .7 confidence

and 10% for .5 confidence. There were no forecasts with less than a .5 confidence assessment selected for investment.

Confidence in the forecasts selected for investment averaged .73. The number of investments per subject ranged from 0 to 13, with a mean of 6.6. On average, subjects wagered $4.23 on each investment; average total investment for each subject was $27.92. Four subjects chose not to invest.

Subjects were more accurate when making decisions (investments) than with judgments. Projection accuracy increased from 29% accuracy for forecasts to 56% when investments were made. The primary reason for this improvement was that subjects chose only the higher confidence forecasts (.5 and up) for investment. In general, accuracy increased with confidence. Thus, by restricting investments to only their relatively higher confidence forecasts, subjects' accuracy improved. This point may seem all too obvious but its significance in decision making may be under-appreciated. It leads to a reduction in the negative consequences of poor judgment from what might otherwise be expected.

A simple regression was run using 'judgmental accuracy' as the independent variable and number of investments made as the dependent variable. Results show that more accurate subjects invested more frequently ($F(1, 23) = 5.65$, $p < .03$, $r2 = .21$). A second regression of 'judgmental accuracy' on level of subjective-risk (operationalized by comparing subject's overall forecast accuracy rate with the confidence assessment of the lowest confidence forecast chosen for investment) indicates that the more accurate subjects assumed greater subjective risks as well ($F(1, 23) = 3.16$, $p = .09$, $r2 = .13$). In other words, the more accurate subjects made investments in forecasts for which they expressed lower confidence assessments when compared with the relatively less accurate subjects. These results suggest that willingness to make decisions varies directly with the competence of the decision maker.

Subjects were also better calibrated when making investment decisions than they were for judgments alone. Figure 1 shows the calibration curves for judgments and investment decisions. The calibration curve for decisions is shifted upwards towards the optimal calibration line from the calibration curve for judgments. A discussion of how this improvement in calibration may have been achieved will be reported later in this paper.

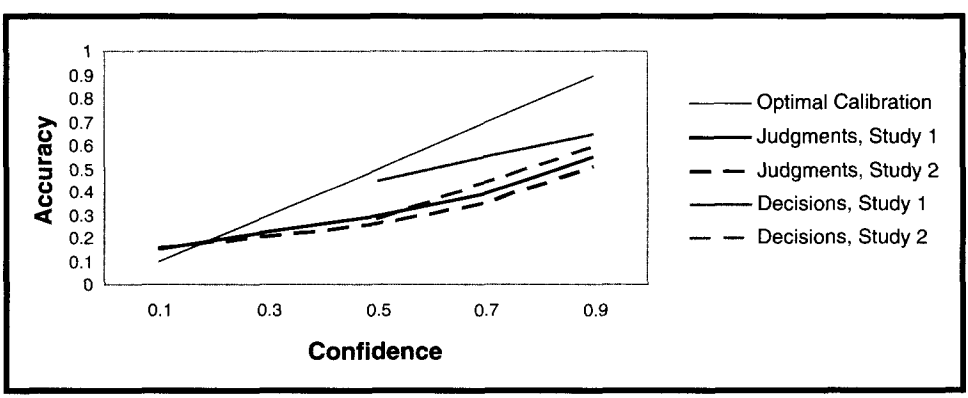

Figure 1. Calibration curves for judgments and investment decisions.

In total, subjects earned $317 on their $210 allocation (twenty-one subjects given $10 each). Earnings ranged from $0 to $33, with mean earnings of $15.10.

STUDY 2

The sample size in Study 1 was somewhat small at 25 subjects. The size did not appear to affect the significance of the results but it does reduce the generalizability of the findings. In addition, subjects in the first study were paid incentives for the accuracy of their forecasts and for selecting accurate forecasts in which to invest but were not offered an incentive for being well calibrated. It is possible that the improvement in decision making over judgment in Study 1 could have been the result of disproportionate incentives. Without a cash incentive, subjects may have taken less care when making confidence assessments, creating doubt as to whether the reported improvements in decision efficacy reported in Study 1 were real, or an outgrowth of the demand characteristics of the study.

A second study was run to address these concerns. Study 2 involved 28 subjects from the same Executive MBA program, in a following year. The population was very similar to those involved in the first study; subjects were from middle and upper management with mean age of 40. Ten of the 28 subjects (37%) were women.

Study 2 was identical to Study 1 except for one addition; cash rewards of $50 and $20 were also offered for the two best calibrated subjects in Study 2. Calibration was explained to subjects and calibration curves were drawn to demonstrate improper calibration. Thus subjects in Study 2 were offered incentives for judgmental accuracy, judgmental calibration and decision making accuracy.

Results

Compared with Study 1, subjects in Study 2 were more overconfident in their judgments. Brier scores were actually higher (worse) in Study 2, in spite of the cash incentives for calibration (Study 1 = .23; Study 2 = .27; $t\ 2.51 = 2.36$; $p < .05$). Average confidence for all forecasts was 51%. Forecast accuracy ranged from 20% to 42% with a mean of 29%.

Subjects in Study 2 made fewer investment choices than those in Study 1. Subjects chose only 135 investments out of a possible 926, or 15% of the total (Study 1 = 26%). The number of investments per subject ranged from 0 to 13, with a mean of 4.9 (Study 1 = 6.6; $t\ 2.51 = 1.65$; $p = .10$). The average investment was $4.64. Thus, on average, subjects invested a total of $22.74.

As in Study 1, subjects in Study 2 were more accurate and better calibrated in decision making than in judgment (Figure 1). Accuracy improved to 53% for investments from 29% for judgments. A regression of 'judgmental accuracy' on 'number of investments made' shows again that the more accurate subjects invested more frequently ($F(1, 26) = 9.7$, $p < .01$, $r2 = .27$). A regression of 'judgmental accuracy' on 'level of subjective risk' in investments shows similar results to Study 1 with the more accurate subjects taking greater risks as well ($F(1, 26) = 3.42$, $p = .07$, $r2 = .12$). Pooling the results from Study 1 and 2 increases the significance of these two regressions substantially ($F(1, 51) = 13.9$, $p < .001$, $r2 = .21$; $F(1, 51) = 6.7$, $p = .01$, $r2$

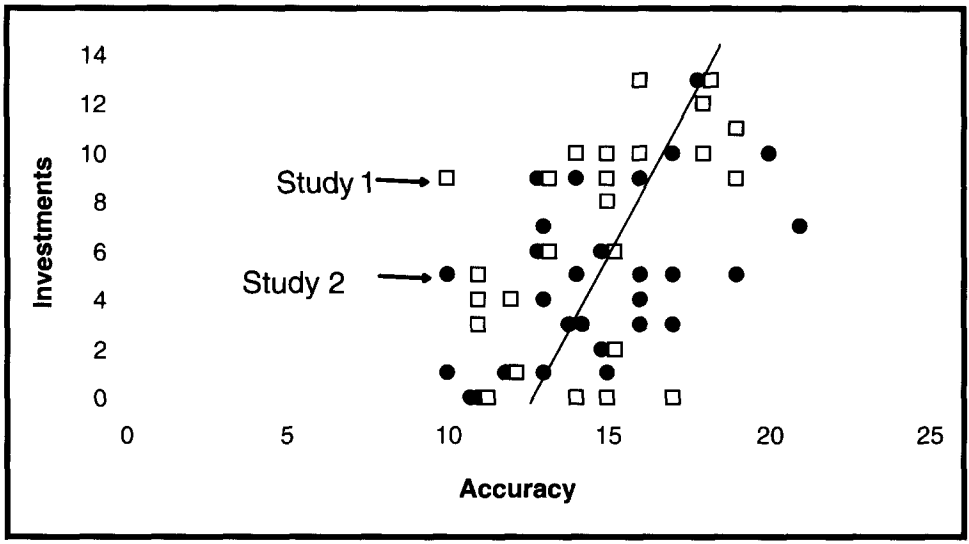

Figure 2. Scatter plot and fitted regression line for the pooled regression of judgmental accuracy on number of investments made.

= .12; respectively). Figure 2 shows the scatter plot and fitted regression line for the pooled regression of 'judgmental accuracy' on 'number of investments made'.

In total, subjects earned $333 on their $250 allocation. Earnings ranged from $0 to $25, with mean earnings of $12.81 (Study 1 = $15.10; t 2, 51 = .75; p = ns).

RATIONALE FOR IMPROVED CALIBRATION

The shift upwards in calibration curves for decisions compared with judgments, seen in both studies, suggests that subjects were better calibrated when making investment decisions than when making forecasts. In other words, this suggests that the act of making a decision led to improved calibration. For this to have happened, subjects had to be able to discriminate between judgments that had a higher chance of winning from those with a lower chance, *at the same level of confidence*, when selecting their investments. Otherwise, calibration would not have been better for investments compared to forecasts. However, it is also possible that the shift in the overall calibration curves may instead have been the result of better calibrated judges taking more frequent action and thus being disproportionately represented in the calibration curves for decisions. This too would shift the calibration curves upward without necessarily implying improved individual calibration brought on by the decision process.

To determine whether calibration improved as a result of the process of making a decision, two additional tests were conducted. The first was designed to determine whether individuals were better calibrated in decision making than in judgment. The second was designed to test for a relationship between calibration and decision frequency.

In the first test, the accuracy rates for forecasts were compared with the accuracy rates for investments, at each level of confidence, for each subject. Specifically, the accuracy rates for

forecasts where no investments were made (judgments only) were compared with the accuracy rates for forecasts where investments were made, at the same level of confidence. With this comparison, a percentage of change, positive or negative, was calculated for each subject. This paired-comparison of judgments and decisions was conducted for all of the subjects pooled from both studies. A sign test was used to determine whether there was a difference in accuracy rates.

For this comparison to be possible, subjects had to forego investments in forecasts in which confidence was either equal to or greater than the confidence of the forecasts in which investments were made. For example, some subjects made investments in .9 confidence level forecasts while skipping other .9 confidence level forecasts. As well, some subjects skipped a forecast with a high level of confidence (.9) while making investments in a forecast with a lower confidence level (.7). Presumably, this occurred because subjects re-assessed their initial judgments when choosing investments, but without making the effort to change their initial recorded judgments. In these cases, a comparison between the accuracy rates for investments and forecasts-without-investments, at the same level of confidence, can be made.

In total, 46 subjects made investments. However, 19 of these subjects employed a simple transference of their highest probability forecasts, to investments. For these subjects, no comparison between forecast accuracy and investment accuracy is possible. (These subjects have an accuracy rate for investments but no corresponding rate for forecasts-without-investments at the same level of confidence.) Typically, this happened when a subject made investments in all of his/her .9 confidence forecasts but made no other investments.

The remaining 27 subjects made investment choices where they either skipped a higher confidence forecast or passed by a forecast of equal confidence. It would thus appear, that for these 27 subjects, an additional evaluation stage occurred between the judgmental stage and the decision making stage. Rather than simply transcribing their highest forecasts for investment, these subjects appear to have made an additional evaluation of their high confidence forecasts which lead them to further discriminate amongst these forecasts. In other words, these subjects used a 'think twice' approach to investing; prior to taking action, subjects reconsidered their initial assessments to ensure that they were well chosen.

Of these 27 subjects, 23 had an accuracy rate for investments which was greater than the accuracy rate for forecasts-without-investments ($p = .0002$, sign test). The median improvement in accuracy was 28%. Thus, on average, the additional evaluation stage significantly improved calibration.

The second test, to determine whether there was a relationship between Brier scores (calibration) and decision frequency, was conducted by running a simple regression with Brier scores as the independent variable and the number of investments made as the dependent variable. No relationship was evident. This is somewhat surprising since intuitively it would seem likely (or perhaps maybe just hoped for) that better calibrated judges would make more frequent decisions.

Taken together, the results of these two tests suggest that the improvement in calibration for decisions compared with judgments occurred on an individual level and was stimulated by

the process of selecting suitable forecasts for investments; it was not the result of better calibrated judges making more frequent decisions. Stated differently, these results suggest that it is not the case that better calibration leads to more frequent action, but rather that taking action leads to better calibration, by stimulating a more thorough evaluation of judgments.

GENERAL DISCUSSION

This paper began by asking whether overconfidence leads to poor decisions. The answer suggested by results from this study is that decisions from overconfident judges may be better than expected. Overconfidence in judgment would be expected to lead to poor decisions if there were no intervening steps between judgment and decision making. However, judgment and decision making are not always one and the same.

Overconfident judges avoid ill-advised action by limiting their action to a subset of their judgments. They tend to be timid. By taking action only when confidence is fairly high, these individuals perform better than might otherwise be expected. This occurs because higher levels of confidence generally correspond with greater accuracy rates. Thus ill advised action is limited, even in the presence of overconfidence.

When the overconfident judge decides to take action, the act of making a decision stimulates the very process which reduces overconfidence. The best way to reduce overconfidence is to get individuals to think of reasons why they might be wrong (Koriat, Lichtenstein, & Fischhoff, 1980). This may be a natural act when moving from judgment to decision making. Results from this study suggest that individuals, when faced with consequences for actions, appear to 'think twice' about their judgments, which leads to improved thoroughness of information processing and better calibration. Thus, overconfidence may be self correcting, to a certain degree, when investigated in decision making (as opposed to the study of judgment only).

Ideally one would like decision makers to be perfectly calibrated. As a second choice, one would probably want the most competent decision makers to take the greatest risks and make the most decisions. Results from this study suggest that relatively more competent decision makers actually make more frequent decisions and choose riskier options than relatively less competent decision makers. In the aggregate, this dampens the impact of overconfidence by having the more accurate decision makers over-represented in the total sample of decisions versus judgments. These results suggest that, in general, individuals have a macro sense of the appropriateness of their knowledge and act accordingly. How does one measure this 'sense' of appropriate knowledge? Earlier studies of confidence have found subjects to be overconfident when dealing with specific items but generally accurate when asked to estimate their overall performance (Kahneman & Lovallo, 1992; May, 1987). Perhaps this overall sense is what governs an individual's willingness to act on judgments. Proper calibration on this more macro question may be relatively more important to the study of decisions as opposed to judgments alone, and should be an area for further investigation.

Care should always be taken when generalizing from results of a particular study. Subjects in this study did improve calibration when dealing with decisions rather than judgments alone,

and in general made effective decisions in spite of being overconfident. It may, perhaps, be reasonable to expect similar results in other settings, particularly when potential consequences to actions loom large. This would be the case in a wide range of decisions which involve financial ramifications, particularly when the ramifications are felt personally, or when notoriety brought on by outside (particularly public) scrutiny is present (Tetlock, 1983; Tetlock & Kim, 1987). However, additional studies should investigate the robustness of these findings and their sensitivity to the nature of actual outcomes in other, specific settings before discounting the concerns raised in the overconfidence literature.

REFERENCES

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

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

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

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

Kahneman, D., & Lovallo, D. (1993). Timid choices and bold forecasts: A cognitive perspective on risk taking. *Management Science, 39* (1), 17-30.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision making under risk. *Econometrica, 47*, 263-291.

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

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

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 in 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and* biases (pp. 306-334). Cambridge: Cambridge University Press.

May, R. S. (1987). *Calibration of subjective probabilities: A cognitive analysis of inference processes in overconfidence*. Frankfurt, Germany: Peter Lang.

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

Paese, P. W., & Feuer, M. A. (1991). Decision, actions and the appropriateness of confidence in knowledge. *Journal of Behavioral Decision Making, 4*, 1-16.

Russo, J. E., & Schoemaker, P. J. H. (1992). Managing overconfidence. *Sloan Management Review, 4*, 7-17.

Tetlock, P. E. (1983). Accountability and complexity of thought. *Journal of Personality and Social Psychology, 45*, 74-83.

Tetlock, P. E., & Kim, J. I. (1987). Accountability and judgment processes in a personality prediction task. *Journal of Personality and Social Psychology, 52*, 700-709.

Tomassini, L. A., Solomon, I., Romney, M. E., & Krogstad, J. L. (1982). Calibration of auditors probabilistic judgment: Some empirical evidence. *Organizational Behavior and Human Performance, 30*, 391-406.

Yates, J. F. (1988). Analyzing the accuracy of probability judgments for multiple events: An extension of the covariance decomposition. *Organizational Behavior and Human Decision Processes, 41*, 281-299.