11-2007

Coherence and Consistency of Investors' Probability Judgments

David V. Budescu
Fordham University (at time of publication: University of Illinois at Urbana-Champaign)

Ning Du
University of Illinois at Urbana-Champaign

MANAGEMENT SCIENCE
Vol. 53, No. 11, November 2007, pp. 1731–1744
ISSN 0025-1909 | EISSN 1526-5501 | 07 | 5311 | 1731

informs

DOI 10.1287/mnsc.1070.0727

Coherence and Consistency of Investors' Probability Judgments

David V. Budescu
Department of Psychology, University of Illinois at Urbana-Champaign, Champaign, Illinois 61820, dbudescu@uiuc.edu

Ning Du
School of Accountancy and Management Information Systems, DePaul University, Chicago, Illinois 60604, ndu1@depaul.edu

This study investigates the quality of direct probability judgments and quantile estimates with a focus on calibration and consistency. The two response modes use different measures of miscalibration, so it is difficult to directly compare their relative (in)accuracy. We employed a more refined within-subject design in which decision makers (DMs) used both response modes to make judgments about a random sample of stocks accompanied by identical information to facilitate comparison between the two judgment methods. DMs judged the probabilities that the stocks will reach a certain threshold, provided lower and upper bounds of these forecasts, and estimated median, 50%, 70%, and 90% confidence intervals of their future prices. We found that the judgments were internally consistent and coherent, but in most cases they were slightly miscalibrated. We used several new methods of analysis that allow for more precise and reliable comparison between the two response modes. We inferred point probability estimates for the target events from the confidence intervals and analyzed them by the same methods applied to binary judgments. Interestingly, when we quantified miscalibration in identical fashion for both methods we did not find evidence of differential levels of miscalibration for the probability judgments and the confidence intervals. We discuss the theoretical and practical implications of these results.

Key words: overconfidence; calibration; confidence intervals; probability judgments; response modes; investment decision; forecasts
History: Accepted by Detlof von Winterfeldt, decision analysis; received August 15, 2005. This paper was with the authors $3\frac{1}{2}$ months for 2 revisions. Published online in *Articles in Advance* October 5, 2007.

1. Introduction

Investing in the stock market is a prototypical case of decision under uncertainty. Investment decisions often rely on probabilistic judgments about future outcomes, and investors are often described as overconfident (Daniel et al. 1998, 2001; De Bondt and Thaler 1995). In the finance literature overconfidence has been associated with real economic consequences of stock return volatility and financial losses (Barberis and Thaler 2003). Overconfidence manifests itself in several ways: Investors may hold unrealistic beliefs about how high their returns will be, or overestimate the precision of their private information and provide too-tight confidence intervals for the future value of stocks (Barberis and Thaler 2003). Extensive empirical evidence shows that overconfident investors take too many risks, trade too much to their detriment, and earn lower average returns (Barber and Odean 2000, 2002). The quality of investors' probabilistic judgments, therefore, influences their choices and affects their investment outcomes.

Recent experimental evidence demonstrates that investors are not uniformly overconfident. Glaser et al. (2003) instructed two groups of investors to predict trend of future price based on past price information. When the investors forecasted future prices via confidence intervals they were overconfident, but underconfidence was observed when they estimated the probability for the price trend. Simultaneous over- and underconfidence was also observed by Kirchler and Maciejovsky (2002) in an experimental market setting: Depending on how confidence was measured, some participants could be classified as either over- or underconfident. The differential degree of overconfidence elicited by these two methods highlights the difficulty of properly assessing judgment quality.

Given the limitations of human attention, memory, and information-processing capacity, it is not surprising that investors' subjective probabilities are often poorly calibrated (Kahneman et al. 1982, Gilovich et al. 2002). Empirical studies documented a general, but not universal, pattern of overconfidence, and showed that the degree of overconfidence often depends on the specific response mode used to elicit subjective probabilities. The literature documents substantial overconfidence in estimates of quantiles (inferred from X% confidence intervals), but lower (or, occasionally, no) overconfidence when people provide

direct probability estimates of binary events (Juslin et al. 1999, 2000; Klayman et al. 1999; Lichtenstein et al. 1982). These two estimation methods have important analogues in the investment setting. For example, investors may decide to buy (or sell) a particular stock only if the probability of the stock price exceeding (falling short of) a certain threshold is X%, i.e., depending on a probability judgment of a binary event. Sometimes, investors rely on a "margin of error" strategy—they buy (or sell) a stock when the price is within a Y% confidence interval.

The purpose of the present study is to examine the quality of confidence judgments regarding future prices of financial assets in a variety of related tasks. This is done in two within-subjects experiments involving a random sample of publicly traded firms. Our results allow us to compare the quality of estimates obtained from the various elicitation methods by focusing on two important features of judgment quality—calibration and consistency—and to shed new light on the underlying sources of the miscalibration typically found in these judgments.

1.1. Different Elicitation Methods and Miscalibration

Extensive research has focused primarily on one facet of judgment quality, calibration—the match between subjective probabilities with the corresponding fraction of actual realizations of the target events (e.g., Budescu et al. 1997b, Gigerenzer et al. 1991, Lichtenstein and Fischhoff 1977, Lichtenstein et al. 1982, Von Winterfeldt and Edwards 1986, Yates 1990). These studies show that people are systematically overconfident about the accuracy of their knowledge and judgments, because their subjective probabilities are frequently more extreme than corresponding accuracy rates. For example, when people express 95% confidence, they may be correct only about 80% of the time. These studies also find that the amount of overconfidence depends on the difficulty of the task. The so-called "hard-easy" effect implies that overconfidence is higher in hard tasks, but attenuated, or even eliminated, in easy tasks (e.g., Lichtenstein and Fischhoff 1977, Lichtenstein et al. 1982, Keren 1991), although recently the reality of this effect was questioned (Juslin et al. 2000). Calibration studies use two types of response modes: estimation of quantiles (sometimes refereed to as fractiles) of probability functions of continuous variables, and probabilistic judgments about discrete propositions (Keren 1991).

Estimates of the quantiles of probability distributions are used for uncertain continuous quantities. Judges are required to provide intervals (values) that correspond to prestated probabilities (Juslin et al. 1999, Keren 1991). Over- or underconfidence is measured by the rate of surprises, i.e., the percentage of true values falling outside the confidence intervals. For example, consider an investor who is asked to provide 90% confidence intervals for a variety of stocks at the end of the year. If the investor is perfectly calibrated, 90% of bounds he or she provided should include the actual values (and 10% of the values should fall outside the stated intervals). If the percentage of surprises is higher than 10%, and the proportion of values in the intervals is lower than the prestated probability (e.g., only 40% of true values fall within the 90% intervals), it is inferred that the judge is overconfident. Conversely, underconfidence is inferred when the proportion of true values in the interval is higher than the prestated probability. The common finding is that the empirical intervals are far too narrow. Hit rates in many studies using 90%–99% confidence intervals are less than 50%, leading to surprise rates of 50% or higher instead of the 1%–10% expected from well-calibrated judges (Alpert and Raiffa 1982, Klayman et al. 1999, Lichtenstein et al. 1982, Seaver et al. 1978).

Direct probability estimates of binary events use the full-range or the half-range assessment method. In the former, judges are asked to assess the probability that various statements are true (or that certain events will occur) on a scale ranging from zero (certainly false) to one (certainly true). For example, one could ask investors to estimate the probability that the stock price for Google will be higher than \$60 at the end of the year (and other similar questions about other stocks). In a half-range task people first decide whether a statement is true, and then assign a probability to this decision. For example, when asked if the price of Google will exceed \$60 per share at the end of the year, they need to agree or disagree with the statement and assess the probability that this choice is correct, using a scale from 0.5 (random choice) and 1 (certainly true). Judges are considered well calibrated if the relative frequencies of true statements match the stated probabilities (e.g., 90% of all events assigned probability 0.9 should be correct). The calibration curve plots the proportion of true (correct) items as a function of the judges' probabilities. The 45-degree line represents perfect calibration, and points below (above) this line reflect over- (under-) confidence (Lichtenstein and Fischhoff 1977). The Brier score and its two components—calibration (or reliability) and resolution—provide quantitative measures of the quality of these judgments (Brier 1950; Murphy 1973; Yates 1982, 1990). Most studies find overconfidence (e.g., Lichtenstein et al. 1982), but conservatism (underconfidence) was also observed (e.g., Edwards 1968, Erev et al. 1994).

Winman et al. (2004) suggested that probability estimates and confidence intervals are formally equivalent because high (low) uncertainties can be expressed

either by low (high) probability judgments or by wide (narrow) interval estimates. Empirically, however, different elicitation methods have produced systematically different judgments (Rottenstreich and Tversky 1997). Although both methods tend to find miscalibration, prior studies suggest that the direct probability judgments induce only a modest bias as compared to the fractile method (e.g., Klayman et al. 1999, Juslin et al. 2000). For example, Klayman et al. (1999) documented less than 5% overconfidence on average when decision makers (DMs) estimated probabilities directly, but documented 45% overconfidence (in 90% confidence interval) when DMs answered confidence-range questions with the fractile method. Some studies using direct probability judgments found modest underconfidence (Erev et al. 1994). Juslin et al. (1999) referred to the pattern of extreme overconfidence with the fractile estimates and the better calibration with the probability estimates as format dependence.

Two main classes of explanations have been offered for overconfidence. These assume either (a) biases in various stages of information processing, or (b) effects of unbiased judgmental error (Soll 1996). Earlier research attributed overconfidence to cognitive biases in information processing and theorized that overconfidence results from biased retrieval and interpretation of evidence (e.g., Hoch 1985, Klayman et al. 1999, Koriat et al. 1980). Other researchers argued that overconfidence is related to unsystematic imperfections in judgment (Budescu et al. 1997a, Erev et al. 1994) because random factors are involved in all stages of the response process, i.e., when people learn the predictive validity of different sources of information (Gigerenzer et al. 1991, Soll 1996), evaluate the available information, and map their subjective feelings of confidence to numerical responses (Erev et al. 1994).[1]

Both errors and biases have been invoked to explain the differential degree of miscalibration associated with different response modes—direct probability estimates and fractile estimates. Winman et al. (2004) attributed format dependence to the naïve use of samples to estimate population properties—specifically, people's tendency to estimate confidence intervals from the sample dispersion they experience, and to estimate probabilities from the sample proportion. Unlike a sample proportion, the sample variance is a biased estimator of population parameter. The failure to correct and adjust for the bias produces too-tight intervals that yield overconfidence (Winman et al. 2004). Alternative explanations of this intriguing pattern focus on two biases in the judgment process—anchoring-and-adjustment and confirmation. According to the anchoring-and-adjustment heuristic, judges start with their single best guess of the quantity (the anchor), and then adjust their estimates. Because their adjustments are insufficient, in general, their interval estimates are too tight (Tversky and Kahneman 1974). Others proposed confirmation bias as a potential source of extreme overconfidence (Klayman et al. 1999, Soll and Klayman 2004). Because the fractile method offers no explicit alternative, people are more likely to form an initial impression of quantity estimates and subsequently retrieve and interpret evidence that confirms these initial estimates (Klayman 1995). Soll and Klayman (2004) argued that, similar to the process of confirming an initial hypothesis, interval estimates tend to be treated as a single (fuzzy) judgment, and the single search of the relevant information creates a narrow range that is often too tight. On the other hand, the response mode of direct probability judgment (half range) offers two explicit alternatives (e.g., the stock is higher, or lower, than a threshold) and encourages people to engage in two separate information searches, one below and one above the given value. According to this explanation, miscalibration of direct probability judgments is more likely to result from random errors rather than cognitive biases.

1.2. The Present Study

The standard finding in the judgment literature (and the generalization that permeates the general literature) is that people's probability judgments are overconfident. However, as discussed in the previous section, overconfidence is far from uniform and universal. In particular, it seems to vary dramatically across elicitation methods. The first, and primary, goal of this study is to directly compare the quality of the judgments obtained from the two methods. There is relatively little work about this important issue, and those few studies comparing them suffer from some glaring deficiencies: (a) they rely almost exclusively on between-subjects comparisons; (b) they do not necessarily use the same items and events; (c) they focus on very general patterns (do judges appear over- or underconfident under both methods?), but do not compare them in a precise manner. In fact, (d) the two methods use different measures of miscalibration and, at this time, there is no good way to compare, directly and accurately, the accuracy of judgments resulting from the use of the two methods. Thus, it is difficult to determine if the observed differences merely reflect different levels of suboptimality,

[1] Some researchers suggested, correctly in our opinion, that overconfidence is, at least in part, an experimentally induced artifact as experimenters often choose harder-than-normal and unrepresentative items (Gigerenzer et al. 1991, Juslin 1994). In this study DMs made judgments regarding future stock prices of real companies that were selected at random. This natural task and the random selection circumvent the problem of unrepresentative sampling often encountered in studies using general knowledge questions selected by the experimenters.

or deeper qualitative differences, and it is impossible to tell whether these discrepancies reflect variability induced by the elicitation methods, the events, or the respondent populations and their differential levels of knowledge and information. The present study seeks to remedy some of these deficiencies and generate data that would allow direct and uncontaminated comparisons between the two methods.

We report results from two experiments. DMs made a series of judgments regarding future stock prices after observing past price performances. Unlike prior studies, which relied on between-subjects designs (Juslin et al. 1999, Klayman et al. 1999, Soll and Klayman 2004), we used a within-subjects design where DMs made six judgments after seeing the same information about 40 real companies. The items used are sampled from a broader universe of relevant events so that they represent an ecologically valid sample, and they are presented to the judges in a fashion that equates the level of relevant information for all participants in all cases. The DMs estimated (a) the probability that the future price will exceed \$20 and (b) the lower and upper bounds (i.e., an interval estimate) of this probability. Although such intervals were rarely used in the context of calibration research, they are quite a common way to model decisions with vague and imprecise information (Budescu et al. 2002, Du and Budescu 2005, Kuhn and Budescu 1996). DMs were also asked to provide (c) 50%, (d) 70%, (e) 90% confidence intervals, and (f) the median price. The second experiment is an extended replication of the first, where we controlled for order effects and introduced more diverse events for probability predictions.

Confidence intervals generated through (implicit) fractile estimates are too narrow (implying large levels of overconfidence), but these studies tend to use only one confidence level. Typically, the focus is on 90% intervals (see, for example, Russo and Schoemaker 1992), or higher confidence levels (e.g., 95%). It is not clear whether people are equally, or even consistently, overconfident across all confidence levels (e.g., Soll et al. 1999, Teigen and Jørgensen 2005). A second goal of our study is to address this issue by eliciting multiple confidence intervals (i.e., with various confidence levels) for the same events from our respondents. This will enable tests of the generality of the overconfidence tendency, as well as new tests of coherence and internal consistency. This will be achieved by comparing the confidence intervals corresponding to three (50%, 70%, and 90%) confidence levels. Given our interest in both methods, we will also study the consistency of the direct probability estimates by comparing precise (point) probability judgments to vague probability judgments. Finally, we will develop new analysis techniques to study, for the first time, the consistency of the two sets of estimates.

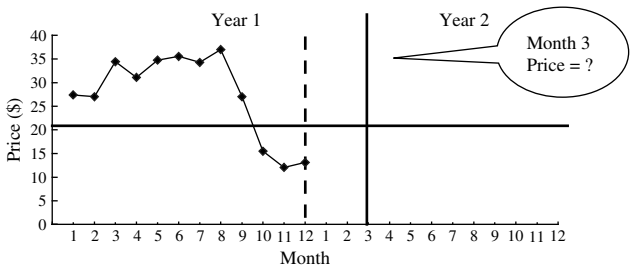
2. Experiment 1

2.1. Experimental Method

Sixty-three graduate accounting students (31 women and 32 men) were recruited in the Business School of the University of Illinois at Urbana-Champaign. They were all familiar with the basic concepts of finance and investment. The participants were asked to assume the role of an investor and provide forecasts for 40 different stocks, randomly selected from a database that tracks stock prices of publicly traded companies. For each stock we presented monthly price series for the 12 months of Year 1. The stocks were not identified by name. We asked DMs to carefully examine each plot and make judgments and predictions regarding prices at the end of Month 3 for Year 2. Figure 1 presents an example of such a price plot. We ran three small groups in class settings involving 30, 16, and 17 DMs. The experiment consisted of two tasks (see details below). The order of the two tasks was randomized in each session. After a brief verbal explanation of the tasks, DMs were given five minutes to read the general instructions. Then we projected the price plots on the screen, one at a time, in random order. DMs wrote down their judgments after each price plot was shown. On average, DMs required 60 minutes to complete the experiment. To motivate the participants we scored their performance according to its accuracy. The scores were calculated by comparing participants' predictions to the true prices for each stock. All DMs were entered in lotteries for cash prizes of \$50. Their chances of winning were proportional to their accuracy scores. Lottery drawings were conducted one week after the experimental session. Four participants were awarded cash prizes.

All participants performed two tasks consisting of six types of judgments. For each judgment we presented the same 40 price plots, but in different random orders. The first task was of probability judgments that required two responses—the best estimate and an interval (range) estimate. The instructions read: "Please indicate how probable it is that the price of each stock will exceed \$20 at the end of Month 3 of Year 2. You must give your best estimate of this

Figure 1 **Monthly Price Plot**



probability (lower and upper bounds of this probability)." The second task asked for estimates of quantiles. The instructions read: "Please estimate the future price for each stock at the end of Month 3 for Year 2. We are interested in a range of possible prices that makes you 50% (or 70% or 90%) sure. When you are 50% (or 70% or 90%) sure, you should expect 50% (or 70% or 90%) of the correct answers to lie somewhere between the upper and lower bounds you specified. In other words, you should provide numbers such that, in your opinion, there is a 50% (or 70% or 90%) chance that the interval includes the correct price, and there is only a 25% (or 15% or 5%) chance that the price will be higher than the upper bound, and a 25% (or 15% or 5%) chance that it will be lower than the lower bound." For the median price, we asked participants to provide a single number that they believe is equally likely that the price will be below or above it.

2.2. Results

We first report results for the probability judgment task, followed by the quantile estimation task, and, finally, results involving both tasks. Within each section we first discuss calibration and then consistency. There were no differences between the two presentation orders, so we present the results of the total sample.

2.2.1. The Probability Judgment Task.

Calibration. DMs estimated the probability that the price will exceed \$20, as well as lower and upper bounds of this probability. We also computed the midrange of this interval, so we have four quantities for each of the 2,520 probability judgments (40 stocks $\times$ 63 participants). We grouped these judgments into 13 categories: 0.00 alone, 0.01–0.09, 0.10–0.19, 0.20–0.29, 0.30–0.39, 0.40–0.49, 0.50 alone, 0.51–0.59, 0.61–0.69, 0.71–0.79, 0.81–0.89, 0.90–0.99, and 1.00 alone (this coding scheme ensures that the three anchor values—0, 0.5, 1—do not bias the other categories), and we counted the relative frequencies of stocks with prices above \$20 in each category. Table 1 shows the distribution for the point estimates. The row labeled "subjective probability" is the average judgment provided by the participants across all items in that category. The modal category was 0.5 (351 responses), and the distribution skews slightly towards the left. This is not surprising, given that 35% of the 40 stocks have true prices above \$20. In the lower part of the table, we also present the mean subjective estimates of the lower and upper bounds and the midrange corresponding to the 13 categories. Table 1 indicates that the best estimates, the midranges, and the upper bounds are all higher than the percentage true.

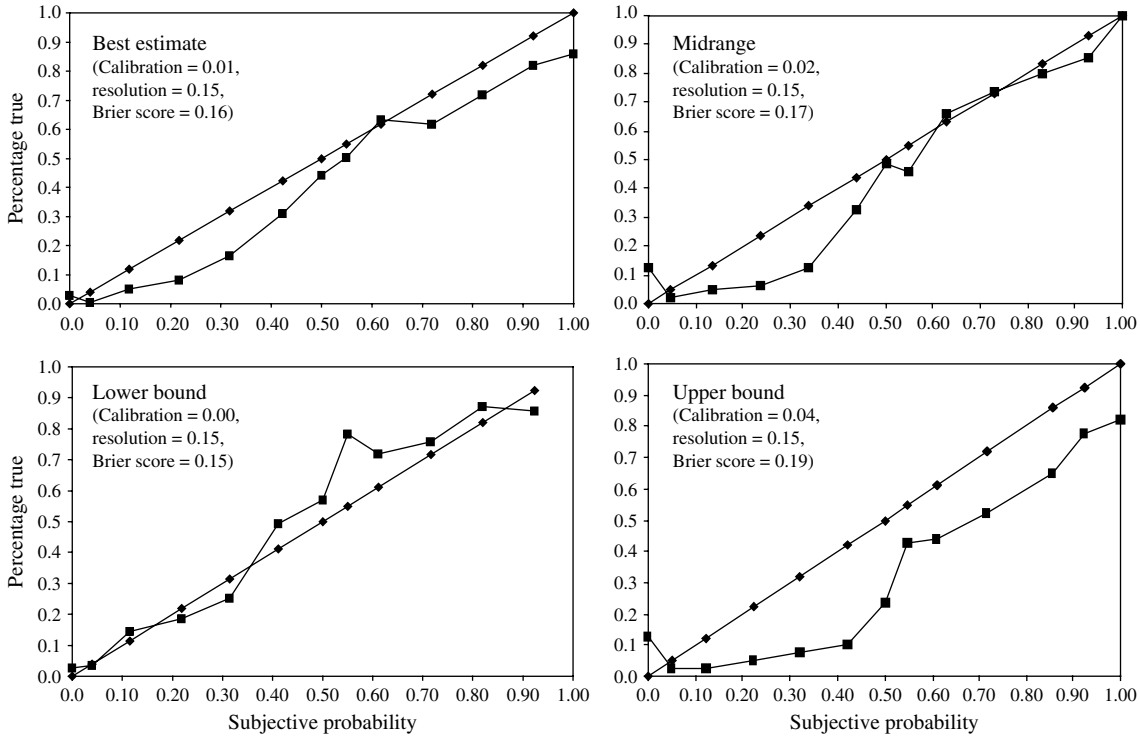
We constructed calibration plots of the relative frequencies of the events, as a function of the mean subjective probabilities for each of the 13 categories. The calibration analysis was performed in four different ways by assigning items to one of the 13 categories based on the best estimate (see Table 1), the lower bound, the upper bound, and the midrange. Figure 2 shows the calibration plots for each of these four groups, along with the Brier score and its two components of calibration and resolution (Brier 1950, 1982). In the best-estimate plot, the subjective probabilities are higher than the observed relative frequencies, indicating a pattern of overconfidence, except for slight underconfidence in the 0 and 0.60–0.69 categories. The calibration plots for the midrange and the upper bound also indicated overconfidence, and the plot based on the lower bound is the most accurate one. The Brier scores do not differ too much from each other, varying from 0.15 to 0.19. Apparently, all four tasks achieve the same level of resolution (0.15), but judgments in the lower-bound category have the lowest (best) calibration score.

Consistency. The first consistency analysis considers proper inclusion: If judgments are consistent, the best estimate should fall within the range defined by the upper and lower bounds, i.e., we expect lower $\leq$ best $\leq$ higher. We found that the majority of judgments—1,657 out of 2,520 (66%)—satisfy this inequality. Violations of proper inclusion are symmetrically distributed—16% of cases have the best estimate lower than the lower bound, whereas 18% of

Table 1 Distribution of Probability Estimates and Their Bounds (Experiment 1)

Categories	0	0.01–0.09	0.1–0.19	0.2–0.29	0.3–0.39	0.4–0.49	0.5	0.51–0.59	0.60–0.69	0.70–0.79	0.80–0.89	0.90–0.99	1
Subjective probability	0.00	0.04	0.12	0.22	0.32	0.42	0.50	0.55	0.62	0.72	0.82	0.92	1.00
Percentage true—best estimate	0.03	0.00	0.05	0.08	0.16	0.31	0.44	0.50	0.63	0.62	0.72	0.82	0.86
Frequency	145	239	260	209	183	265	351	119	232	188	146	133	50
Subjective probability—lower bound	0.05	0.09	0.12	0.18	0.26	0.33	0.37	0.41	0.46	0.48	0.52	0.63	0.65
Subjective probability—upper bound	0.19	0.24	0.32	0.39	0.49	0.56	0.61	0.63	0.68	0.71	0.75	0.82	0.87
Subjective probability—midrange	0.12	0.16	0.22	0.28	0.38	0.45	0.49	0.52	0.57	0.6	0.64	0.73	0.76

Figure 2 Calibration Plots (Experiment 1)



cases have the best estimate higher than the upper bound. The consistency rate increases markedly (to 75%) when we use removed judgments related to the five stocks and the nine participants with the highest rates of violations. We analyzed the consistency rate at the individual level by calculating the mean of each subject's judgments across all 40 stocks. Only 6 out of 63 participants violated this inequality at the aggregate level, yielding a consistency rate of 95%.

We compared the best, the lower, and the upper probability judgments in a one-factor repeated-measures ANOVA. The means of these three estimates are significantly different ($F_{(2,124)} = 201.40, p < 0.01$). The mean of the best estimates is bounded by the means of the lower and upper bounds, as expected. The midrange of the vague interval judgment (i.e., (lower bound + upper bound)/2) is 0.42, and matches the best estimate. Thus, the precise and the vague judgments elicited by the direct probability estimate are highly consistent.

Lastly, we correlated the mean estimates of the best, the lower bound, the upper bound, and the midrange for each of the 40 stocks (across 63 participants). All correlations are higher than 0.90 and significant. We also correlated these measures with the mean width of the range. These correlations are also positive (between 0.38 and 0.41) and significant. The positive relationship implies that participants injected higher uncertainty into their judgments about stocks that were judged more likely to exceed \$20. Apparently, participants hedge their higher-probability judgments by widening the range.

2.2.2. The Quantile Estimates.

Calibration. Participants provided 50%, 70%, and 90% confidence intervals. Judgments are considered calibrated if the fraction of cases that were in fact correct matches the interval's stated confidence. If the confidence is higher (lower) than the actual hit rate, the DMs are said to be overconfident (underconfident). The empirical results (see Table 2) show an interesting pattern of underconfidence at the 50% level, perfect calibration at the 70% level, and overconfidence at the 90% level. Evidently, overconfidence is not universal.

Consistency. DMs provided seven quantiles. If these estimates are fully consistent, they should be ranked accordingly (lower bounds of 90%, 70%, 50%, median, upper bounds of 50%, 70%, and 90%). We calculated the (Kendall) rank-order correlations between these seven estimates and their expected order (it should

Table 2 Calibration of Confidence Intervals (CIs) (Experiment 1)

	50% CI	70% CI	90% CI
Percentage of prices in the interval (hits)	0.59	0.70	0.82
Expected percentage of prices in the interval	0.50	0.70	0.90
Over/underconfidence (surprises)	−0.09	0.00	0.08

be one for perfectly consistent ratings). For each subject, we calculated the means of the seven estimates across all 40 stocks and used these in the calculations. This generated a distribution of 63 individual DM correlations. Kendall's τ_b indicates that DMs are highly consistent: The mean correlation is 0.90 (only one correlation is lower than 0.50) and 62 of the 63 correlations are significantly greater than 0 ($p < 0.05$). For each stock, we also calculated the means of these seven estimates across all 63 DMs. This generates a distribution of 40 individual stock correlations. Kendall's τ_b indicate that price estimates are highly consistent. For 39 stocks, the seven estimates and the expected order are perfectly (ordinally) correlated. Only one stock has a less than perfect correlation coefficient (0.91).

Next, we conducted a two-factor repeated measures ANOVA of the bounds of the confidence intervals. The first factor is the nature of the bound (lower or upper) of the interval, and the second factor is the interval's level of confidence. Naturally, the upper bounds are higher than the lower bounds ($F_{(1,62)} = 423.42, p < 0.01$), and there is a significant effect of confidence level ($F_{(2,124)} = 9.80, p < 0.01$). As expected, we found a significant interaction, reflected in a fanning out pattern ($F_{(2,124)} = 80.77, p < 0.01$) depicted in Figure 3. The line of the midrange is almost straight (it was not included in the analysis, but it is plotted to facilitate interpretation), and bracketed by the lower and the upper estimates. The mean of the median estimates is 18.69 (with a minimum of 17.23, a maximum of 21.65, and a SD of 84), and closely matches the midranges of the confidence intervals. It appears that the DMs' confidence intervals are internally consistent.

2.2.3. Comparing the Probability and the Fractile Methods.

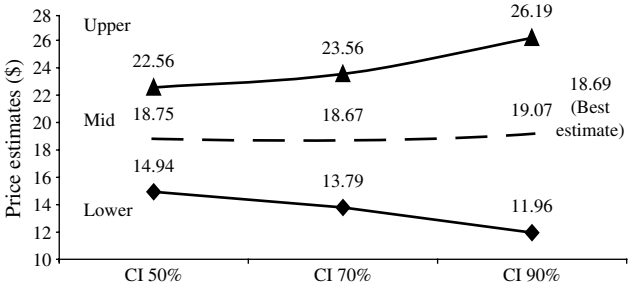
Calibration. If DMs' judgments are consistent across these two elicitation methods, they should exhibit similar patterns of over- or underconfidence. We used a (new, to the best of our knowledge) matching procedure to compare the two sets of judgments by mapping the results from the fractile method to the probability method. We used the seven judged quantiles (Q05, Q15, Q25, Q50, Q75, Q85, and Q95) to estimate the means and standard deviations of the best-fitting normal distribution for each of the 2,520 cases (63 DMs × 40 stocks).[2] These distributions were used to calculate the z scores corresponding to the target price ($20), and the expected probability of the stock being above it. We grouped these 2,520 predicted probabilities into the 13 categories used in §3.1. Table 3 shows the distribution of expected probabilities. Remarkably, we see a very similar pattern when we plot the calibration for predicted probabilities (see Panel B in Figure 4) next to the judged probabilities (Panel A in Figure 4, the same as the best estimate in Figure 2). Both plots show a predominant pattern of overconfidence and almost identical measures of calibration and resolution.

This analysis demonstrates that when judgments elicited by the two methods are evaluated by similar measures they are highly consistent. Moreover, when we used the probability function implied by the quantile estimates to predict the lower and upper bounds of 50%, 70%, and 90% confidence intervals, we found that these predicted values are better calibrated than the empirical data. Specifically, 55%, 71%, and 87% of the true prices fall within the predicted (50%, 70%, and 90%, respectively) confidence intervals (compare with the values from Table 2).

Consistency. First, we test whether the direct probabilities associated with the target event (price ≥$20) are consistent with the boundaries of the intervals in which $20 is located. We inspected each judgment set and determined the location of the target price ($20). For example, if the value of $20 falls between Q15 and Q25, we expect $(1 - 0.25) = 0.75 < (P(X > 20) < (1 - 0.15) = 0.85$ to hold. The expected pattern was found in 1,662 out of the 2,520 cases yielding a 66% consistency rate. The observed violations are almost symmetrically distributed, and the consistency rate increases to 70% when we exclude the nine DMs with the worst performance.

Finally, we compared the probabilities predicted from the fractile method (using the approximations described earlier in this section) with the actual probability judgments. We calculated the difference between the judged and predicted probabilities, and

Figure 3 **Mean Bounds of Different Confidence Intervals (Experiment 1)**



[2] To fit a normal distribution to these estimates, we use the Z values under a normal distribution corresponding to the seven estimates. By definition: $Z = (\text{Price} - \text{Mean})/\text{SD}$. We rearrange terms and get $\text{Price} = \text{Mean} + \text{SD} * Z$. We regressed the judged prices on the seven Z values, so the intercept estimates the mean price of the stock and the slope estimates the SD of the distribution. The expected probability is obtained by calculating the z score corresponding to each stock and subject, and is subsequently determined by using the normal table.

Table 3 Distribution of Predicted Probability Judgments (Experiment 1)

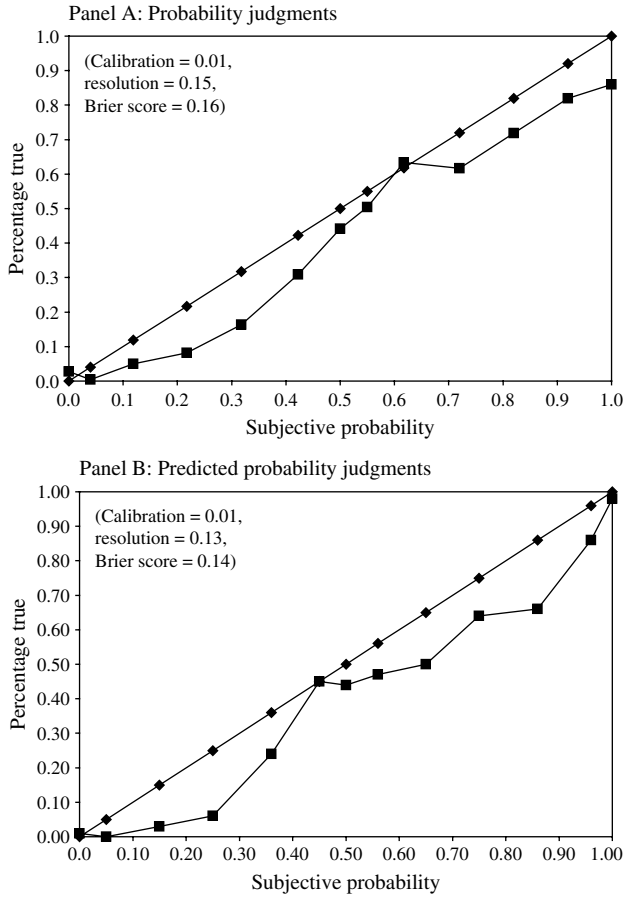
Categories	0	0.01–0.09	0.1–0.19	0.2–0.29	0.3–0.39	0.4–0.49	0.5	0.51–0.59	0.60–0.69	0.70–0.79	0.80–0.89	0.90–0.99	1
Subjective probability—best estimate	0.00	0.05	0.15	0.25	0.36	0.45	0.50	0.56	0.65	0.75	0.86	0.96	1.00
Percentage true	0.01	0.00	0.03	0.06	0.24	0.45	0.44	0.47	0.50	0.64	0.66	0.86	0.98
Frequency	239	407	223	156	138	155	45	256	259	192	144	222	84

conducted *t*-tests to see whether their means are different from zero for each subject across these 40 stocks. Within-subject consistency across methods is very high: *t*-tests found that only 9 persons out of 63 (14%) had mean differences significantly different from zero ($\alpha = 0.0008$ using the Sidack adjustment). Figure 5 presents the mean difference of these two for the best estimate, the lower bound, and the upper bound for each subject. The line for the best estimate is mostly around zero, and is bracketed by the lower and upper bounds.

3. Experiment 2

The most surprising result of the first study was the high level of consistency between the two methods when the results are analyzed by similar methods in a common metric (see Figure 4). One possible explanation is that because the subjects saw the same items several times, they simply recalled some of their earlier judgments and tailored subsequent judgments accordingly. A second study was conducted to replicate and validate the results of the first experiment. The second experiment uses new items, and more events for probability predictions (not only \$20). Also, to refute the possibility that the remarkable level of consistency between tasks observed in the first experiment was due to the blocking of the tasks, we manipulated the order of tasks administered.

Figure 4 Comparison of Calibration Curves (Experiment 1)



3.1. Experimental Method

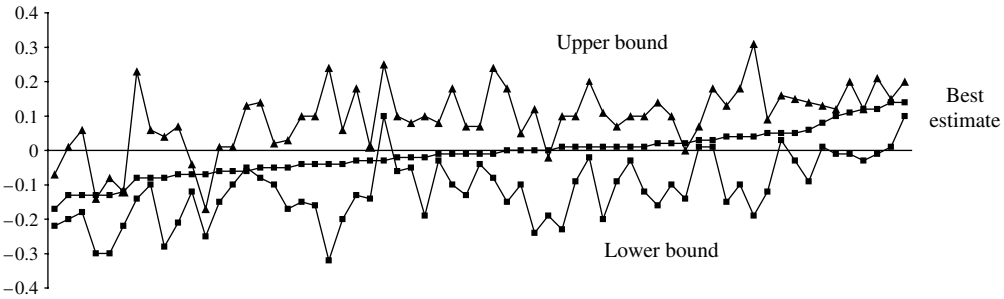
Seventy-five undergraduate business students (39 women and 36 men) from DePaul University participated in this experiment. The study was similar to the first one in most respects, but differed in some details: (a) For each stock, the participants saw time series of the past 24 months (instead of 12 months) and were asked to make forecasts for stock prices at the end of Month 3 of Year 3 (instead of Month 3 of Year 2). (b) The experiment consisted of two tasks—probability judgments and confidence intervals. The confidence interval task was identical to the first study, but we only asked for the best-estimate probability judgment (dropping the request for lower and upper bounds). On the other hand, we asked participants to judge three events: how probable it is that the stock price will exceed \$20, \$30, and \$40, respectively. Thus, participants made seven different judgments—50% confidence interval (CI 50), 70% confidence interval (CI 70), 90% confidence interval (CI 90), median, probability for price \$20, probability for price \$30, and probability for price \$40. (c) We used a different set of (randomly selected from the same database) 40 stocks. (d) Most important, we ran four small groups in class settings involving 19, 18, 20, and 18 DMs. For each group, we used a different order of the seven tasks (see Table 4).

We instructed subjects to be as accurate as possible, and promised financial rewards for accuracy. The two subjects with the highest accuracy scores in each group received \$25 gift cards redeemable at the campus bookstore.

3.2. Results

This section is structured as in the first experiment, but some details are omitted to save space.

Figure 5 Mean Difference Between Probability Judgments and Predicted Probabilities (Experiment 1)



3.2.1. The Probability Judgment Task. In this task DMs estimated the probability that prices will exceed \$20, \$30, and \$40, so we have three quantities for each of the 3,000 probability judgments (40 stocks for 75 participants). If judgments are consistent, we expect the probability judgments to be monotonically increasing (i.e., $(P(X \geq 20) \geq (P(X \geq 30) \geq (P(X \geq 40))$). Sixty-nine of the 75 participants (92%) made probability judgments satisfying this inequality, indicating high levels of consistency.

Panel A of Figure 6 shows the calibration plot (based on 13 categories), as well as the Brier score and its decomposition. The subjective probabilities are lower than the observed relative frequencies, indicating a pattern of underconfidence. Detailed inspection suggests this pattern is driven by the results for \$20 and \$30 price classes, whose true frequencies are 93% and 80%, respectively, whereas the judgments for \$40 were well calibrated. Figure 7 shows the calibration plots of probability judgments for each group. The calibration scores and the calibration plots are very similar for all four groups. Therefore, we conclude that the results were not driven by the different task order.

3.2.2. The Quantile Estimates. The participants provided a total of 9,000 confidence intervals (40 stocks × 75 participants × 3 CIs). We conducted a consistency analysis to determine whether the seven quantiles were ranked appropriately. The mean (Kendall's τ_b) correlation is 0.82 and 74 of 75 correlations are significantly greater than zero ($p < 0.05$), indicating that the DMs are highly consistent.

Table 4 Task Order (Experiment 2)

Task order	Group 1	Group 2	Group 3	Group 4
1	CI 70	CI 90	Price 30	Price 40
2	CI 50	CI 50	Price 20	Price 20
3	CI 90	CI 70	Price 40	Price 30
4	Median	Median	CI 70	CI 90
5	Price 30	Price 40	CI 50	CI 50
6	Price 20	Price 20	CI 90	CI 70
7	Price 40	Price 30	Median	Median
	$N = 19$	$N = 18$	$N = 20$	$N = 18$

The results of the calibration analysis (see the top row in Table 5) were very similar to the first study: underconfidence at the 50% level, good calibration at the 70% level, and overconfidence at the 90%. The last four rows in Table 5 indicate that these rates were very similar for all four orders.

3.2.3. Comparing the Probability and the Fractile Methods. First, we map the results from the fractile method to binary probabilities by estimating the means and standard deviations of the best-fitting normal distribution for each of the 3,000 cases (75 DMs × 40 stocks). These distributions were used to calculate the z scores corresponding to the target price

Figure 6 Comparison of Calibration Curves (Experiment 2)

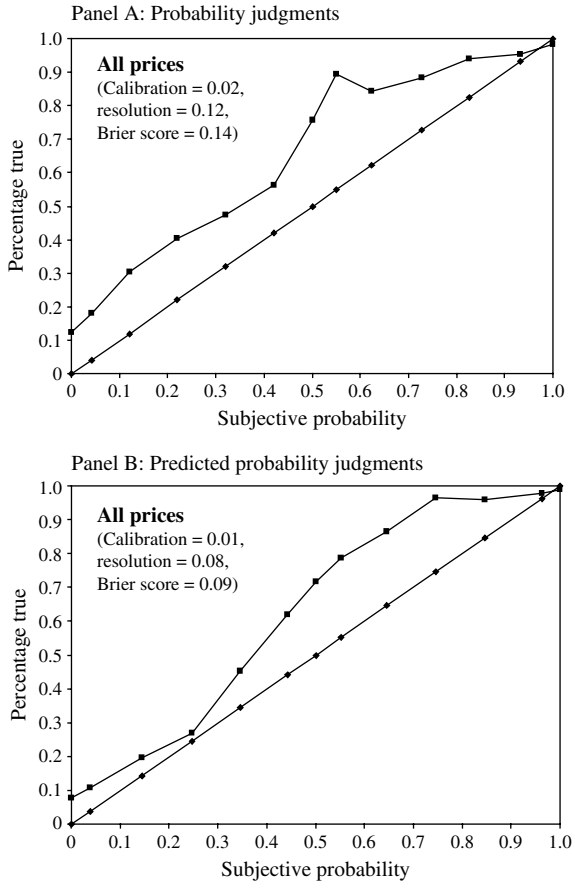
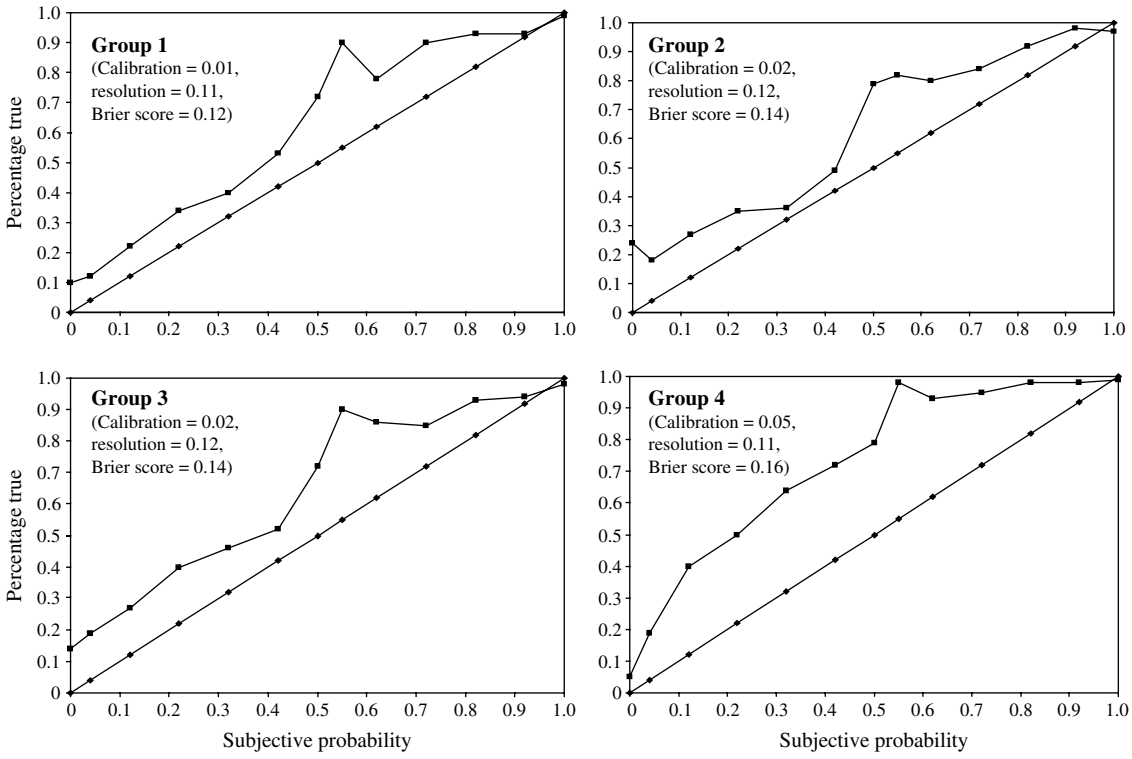


Figure 7 Calibration Plot by Each Group (Experiment 2)



($20, $30, and $40, respectively), and the probability of the stock being above it. For a vast majority of the subjects, the predicted probabilities were not significantly different from the actual judged probabilities. The calibration plot of the 9,000 predicted probabilities (see Panel B in Figure 6) is very similar to the one based on the direct probability judgments (Panel A in Figure 6). Both show a predominant pattern of underconfidence, and have almost identical measures of calibration and resolution.

Next, we tested whether the location of the direct probabilities associated with the target event (price $\geq$$20, $30, or $40, respectively) are consistent with the boundaries of the intervals in which $20, $30, or $40 is located (see the consistency analysis in §2.2.3). The expected pattern was found in 4,815 out of the 8,966, for a 54% consistency rate. If instead of the seven original quantiles we use only five (ignoring the 90% CI), or three (ignoring the 90% and 70% CIs), the consistency rate increases to 62% and 68%, respectively.

Table 5 Rate of Surprises Overall and by Group (Experiment 2)

	50% CI	70% CI	90% CI
All Groups	−0.14	0.01	0.12
Group 1	−0.13	0.06	0.20
Group 2	−0.13	−0.05	0.09
Group 3	−0.12	0.01	0.14
Group 4	−0.18	0.00	0.06

In the present study subjects provided three probability judgments for every stock. We used these probabilities to estimate the means and standard deviations of the best-fitting normal distribution for each of the 3,000 cases (75 DMs $\times$ 40 stocks).[3] Thus, we had two independent sets of estimates of these parameters—one based on the quantile estimates and one based on the probability judgments. The final analysis compared these distinct sets of estimates. First we compared the means of the distributions. We calculated the mean difference for each subject across all the stocks. The mean difference is −0.77, and it is not significantly different from 0 ($t_{70} = -1.24; p > 0.05$). We also calculated the mean difference for each stock across all subjects, and obtained similar results (Mean $= -1.12$, $t_{39} = -1.49; p > 0.05$). Thus, the quantile and probability judgments seem to reflect probability distributions with identical means.

Next we analyzed the ratio of the two standard deviations. First we combined the ratios for each of the subjects across all the stocks. The standard deviations inferred from the probability judgments are higher (median within-subject ratio = 1.67). A test based on the logarithms of these ratios indicates that this difference is significant ($t_{70} = 5.40; p < 0.05$). We

[3] Because we only have three points, it was not possible to obtain meaningful estimates of the two parameters in almost 24% of the cases. This includes cases with incomplete data, equal probabilities, and inconsistent judgments.

performed the same analysis based on within-stock (across all subjects) ratios and obtained similar results (Median ratio $= 2.01$, $t_{39} = 11.03$; $p < 0.05$). The probability judgments seem to reflect probability distributions with higher variances.

4. General Discussion

4.1. Summary of Major Findings

We studied the quality of direct probability judgments and quantile estimates with a focus on calibration and consistency. We employed a more refined within-subject design, in which DMs used both methods to judge random samples of stocks accompanied by identical information to facilitate comparison between the two judgment methods. After observing past performance of a given stock, DMs estimated the probability that the future price will exceed a target price, lower and upper bounds (i.e., an interval estimate) of this probability, and various quantiles of the distribution of future prices.

We found strong evidence that judgments are coherent and consistent within each method. For direct probability estimates, most point estimates fell within the range implied by the lower and upper bounds. The DMs' judgments of the quantiles were sensitive to the degree of confidence, in the sense that they widened their range estimates according to the confidence level. This sensitivity of the interval estimates to the confidence probability level is consistent with some prior findings (Alpert and Raiffa 1982, Juslin et al. 1999, Seaver et al. 1978), but contradicts some recent results. Jørgensen et al. (2004) and Teigen and Jørgensen (2005) found that different groups of subjects provided almost identical confidence intervals for 50%, 75%, 90%, and 99% levels. Several factors can account for the different patterns of results, but the most important is the nature of the design and elicitation procedure. Sensitivity to the confidence levels was found only in studies employing within-subjects designs (like the present one). When DMs provide multiple confidence intervals (e.g., 50%, 70%, and 90%) they can adjust and tailor the widths of confidence intervals according to the whole sequence of judgments. This pattern is not necessarily replicated in between-subjects designs (e.g., in Teigen and Jørgensen 2005), where such direct comparisons between various levels of confidence are impossible.

Moreover, we found strong evidence that judgments are consistent across response modes because most between-methods comparisons showed a high level of agreement. The majority (66%) of the direct probability judgments (see Experiment 1) fell within the interval predicted by the inferred price distribution. Remarkably, the internal consistency of the direct probability judgments (point probability within the range) was also 66%, indicating that the between-methods agreement was as high as one could reasonably expect given the inaccuracy of the methods. In the same spirit, when we compared the probabilities inferred from the confidence intervals with the direct probability judgments, we did not find them to differ significantly. In Experiment 2, we used a different order of the judgment tasks in the various groups, yet we replicated all the major results pertaining both to within-method coherence and between-method consistency. This invariance indicates that the original results were not due to the structure of the original experiment.

Consistent with previous findings that DMs are typically miscalibrated, the direct probability estimates of the binary events displayed modest overconfidence (Experiment 1) and underconfidence (Experiment 2), because the point probabilities did not precisely match the relative frequencies of true events. Judgments elicited by the fractile method showed a more complex and intriguing pattern: too-wide 50% confidence intervals, overly narrow 90% confidence intervals, and perfectly calibrated 70% confidence intervals, in both studies. Unlike prior studies, where 90% confidence intervals yielded very low hit rates (in the 25%–40% range in Soll and Klayman 2004 and Teigen and Jørgensen 2005), hit rates in our experiments are about 80%. We attribute part of the differences to the nature of our stimuli. Many studies documenting extreme overconfidence used general knowledge questions, where people had to rely on, and sample from, their memory, and therefore were more susceptible to cognitive biases (Kahneman et al. 1982, Klayman et al. 1999, Soll and Klayman 2004). Confidence judgments for general knowledge items reflect internal uncertainty where DMs are uncertain about their level of knowledge. They induce large individual differences due to differential levels of information. In our study, DMs predicted future events in the market where a good portion of the uncertainty is due to external sources (fluctuations of the financial market). In fact, Teigen and Jørgensen (2005) have argued that interval estimates make less sense for judgments that probe exclusively internal uncertainty. An additional factor that explains the better quality of judgments in our study is the DMs' access to relevant (past) predictors from which they could infer the future. Prior studies in judgmental forecasting document that serially correlated cues, such as those in time series, tend to reduce the degree of overconfidence when DMs estimate confidence intervals (Lawrence et al. 2006). The availability of these cues allowed DMs to estimate reasonable confidence intervals and considerably reduced (eliminated in some cases) the level of overconfidence (Teigen and Jørgensen 2005).

The differential pattern found in the three confidence intervals can be explained by the trade-off between the two competing objectives—accuracy and informativeness—that people consider when they estimate uncertain quantities (Yaniv and Foster 1995). Wider intervals are less informative, but are more accurate (Yaniv and Foster 1995, 1997). The overreaching goal of the DMs in our study (and the incentive structure we provided) was to maximize the rate of correct judgments, but it appears that this was operationalized in different ways in the various tasks. When asked to provide 50% confidence interval, DMs chose to increase accuracy (i.e., hit rates) by widening their estimated intervals, which induced underconfidence. On the other hand, when they were asked to provide 90% confidence intervals the DMs deliberately avoided wider ranges to ensure that their expressed estimates were sufficiently informative, leading to the observed overconfidence. The 70% confidence intervals seem to be those where most DMs achieved the best compromise between the needs to be informative and to be accurate, i.e., one that produced well-calibrated judgments. Regardless of the validity of this interpretation, our results illustrate the pitfalls of generalizing about, supposedly, global patterns (such as overconfidence) based on limited data (say, only 90% confidence intervals).

The two response modes use different scales and their standard analysis invokes different measures of (mis)calibration. We developed several new approaches to achieve more precise and direct comparison between them. In the most direct and informative comparison, we used the confidence intervals to fit a distribution of future prizes, and predicted the probabilities of the target events (price greater than a given $ threshold) from the fitted distribution. The calibration analysis of these predicted probabilities yielded results (calibration curves and Brier scores) that were very similar to those from the direct probability estimates. In other words, when we used an approach that quantifies miscalibration in identical fashion for both methods, we did not find evidence of differential levels of miscalibration for direct probability judgments and confidence intervals.

4.2. Implications of the Results

Our results confirm some of the standard findings in the probability judgment literature, but challenge some of the generalizations regarding overconfidence that permeate this literature and highlight the need to study the quality of such judgments using multiple elicitation methods. For direct probability estimates, we found only slight miscalibration (overconfidence in one study and underconfidence in the other). This supports the view that when events/items are selected at random from meaningful reference classes, DMs are quite accurate (Gigernezer et al. 1991, Juslin 1993), although not perfect (Budescu et al. 1997a). We also replicated the standard, and often cited, finding of unrealistically narrow 90% confidence intervals. Unlike prior studies, we elicited intervals for three different confidence levels (50%, 70%, and 90%), and found that they are not equally miscalibrated. In fact, the 70% intervals are quite accurate.

The 90% confidence intervals have become the standard tool for measuring calibration, but our findings illustrate quite dramatically the danger of relying on this single measure and extrapolating from it, and infer global overconfidence. In fact, the choice of a specific confidence level should depend on the issues and risks involved (Russo and Shoemaker 1992). For example, if an investor plans to invest a large sum of money in a small high-tech firm, where he/she may face the risk of losing all his investment, the DM may want to incorporate extreme swings in stock prices, and assess a 90% (or higher) confidence level on future prices. Lower confidence levels (e.g., 50% or lower) are appropriate if an investor faces minimal downside risks, but values high accuracy. We observed the least amount of bias when DMs provided 70% intervals. Given this pattern, it may be useful to adopt the use of the 70% intervals in most practical applications to obtain the most accurate information from advisors. We cannot state with confidence that this pattern holds in all contexts and domains, but our results suggest that it is relatively easy to identify those confidence levels that are most accurate (because they achieve the right balance between perceived accuracy and informativeness) for distinct domains. Finally, if one is interested in obtaining a complete picture of the DMs' performance, our study documents the benefits of eliciting multiple intervals involving various confidence levels.

Our results are consistent with the growing body of literature on format effects in judgment, and support the view that confidence interval judgments are sensitive to the specific elicitation method. The degree of overconfidence depends on (a) whether point estimates are made before stating the interval, as opposed to afterward (Juslin et al. 1999); (b) whether DMs answer questions in two-point format with only the boundary points or three-point format with the boundary points along with the median estimate (Soll and Klayman 2004); or (c) which specific prestated probability level (e.g., 50%, 70%, or 90% as in this study) is given for the corresponding intervals (values). And, of course, our results illustrate the differences between the direct and interval judgments. Despite the presence of miscalibration in both methods, we showed that they are quite consistent with each other and, in particular, that one can use estimates of the quantiles to infer quite accurately the

judges' direct probability judgments. This realization has both theoretical and practical implications.

Miscalibration of probability judgments is attributed to systematic cognitive biases and/or the "noise" (the stochastic component) associated with the judgment process (Budescu et al. 1997a, b; Erev et al. 1994; Juslin et al. 1999). Of course, these sources can operate in different ways under the two methods. Our results cannot rule out any source, nor can we determine their relative contribution to overt overconfidence, but we can offer some speculations on the nature of the problem. The most important conclusion stems from the observed agreement between the two response modes and the possibility of using the confidence intervals to predict the judges' direct judgments. This result implies that in both methods DMs rely on common covert subjective feelings of confidence inferred from the information available to them, which are then converted into the overt numerical responses (Erev et al. 1994, Soll 1996) as required by the methods. We cannot conclude that this common internal representation is accurate (well calibrated), but it makes good sense to assume that, if it is biased, the nature, direction, and magnitude of the bias is not method specific and it affects responses in both methods in the same way. The most direct support for this conclusion is the similarity of the mean values of the distributions inferred for the probability judgments and the quantile estimates.

On the other hand, it is reasonable to assume that the error magnitude and effects are method specific because of the different conditioning of the problem in the two tasks. When asked to judge probabilities, the DM conditions the judgments on given values (i.e., judge $F(X)$ given X) and responds on the bounded $[0, 1]$ probability scale. When asked to estimate quantiles, the DM conditions the judgments on fixed probabilities (i.e., estimate X given $F(X)$), and responds on an unbounded monetary scale (price ≥ 0). It is well documented (Erev et al. 1994, Budescu et al. 1997a) that in the first case the errors are regressive towards the mean (i.e., pulling the judgments towards 0.5) because of the boundedness of the scale. There is no reason to expect a similar pattern for quantile judgments but, unfortunately, there are no (and we did not collect) data to quantify these distributions. We do have indirect support for this speculation in the different estimates of variance obtained from the two tasks in the second study. Future work should replicate our study but should replicate (at least some of the) judgments to allow for estimation of within-judge variance for both methods.

Proper assessment of judgment quality is extremely important in investment setting because investors make critical decisions regarding future values of risky assets based on imperfect knowledge and incomplete information. To make sound investment decisions, investors must have a realistic view of the quality of their information. The standard view in the finance literature is that investors are overconfident about their ability to evaluate securities and pick stocks (Barber and Odean 2000; Daniel et al. 1998, 2001; De Bondt and Thaler 1995). Recent experimental results challenged the generality of the overconfidence claim and questioned whether financial models should assume overconfidence when predicting investors' behaviors, or whether they should be context specific (i.e., incorporate either over- or underconfidence depending on the response mode) to improve their predictive validity. Financial researchers are faced with the critical task of choosing a specific method to elicit accurate confidence judgments to minimize the problems associated with miscalibration and to extract valid information from the noisy data.

Thus, it is natural to ask which method(s) one should use to obtain accurate and well-calibrated probabilities regarding financial items. Our recommendation is straightforward: We favor elicitation of multiple quantiles that can be used to fit a complete distribution of the target quantity. This distribution can be used (a) to predict probabilities of specific binary events (e.g., various thresholds), (b) to improve the quality of the target confidence intervals (note that in our study the three confidence intervals inferred from the fitted distributions were more accurate than those based on the judges' raw estimates), and (c) to infer confidence intervals (e.g., 60%, 95%) that were not elicited directly. We recognize that there are many unanswered practical questions, such as how many, and which, quantiles to estimate, what constrains to impose on the fitted distributions, and how to elicit judgments to maximize the accuracy of the distribution. For example, Soll and Klayman (2004) suggested that the precision of the distributions is improved if, instead of judging X% confidence intervals, judges are asked to estimate its endpoints separately (i.e., the $(100 - X)/2$ and the $(100 + X)/2$ quantiles). We hope that these issues will be addressed by future research.

Acknowledgments

This work was supported by grants from the U.S. National Science Foundation under Awards NSF SES 02-41434 and 06-20008. The second author gratefully acknowledges the financial support of the Richard D. and Anne Marie Irwin Foundation at University of Illinois at Urbana-Champaign. The authors thank two anonymous reviewers for helpful comments.

References

Alpert, M., H. Raiffa. 1982. A progress report on the training of probability advisors. D. Kahneman, P. Slovic, A. Tversky, eds. *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press, Cambridge, UK, 294–305.

Barber, B., T. Odean. 2000. Trading is hazardous to your wealth: The common stock investment performance of individual investors. *J. Finance* **55** 773–806.

Barber, B., T. Odean. 2002. Online investors: Do the slow die first? *Rev. Financial Stud.* **15** 455–487.

Barberis, N., R. Thaler. 2003. A survey of behavioral finance. G. M. Constantinides, M. Harris, R. Stulz, eds. *Handbook of the Economics of Finance*. Elsevier, North-Holland, Amsterdam, 1052–1090.

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

Budescu, D. V., T. S. Wallsten, W. T. Au. 1997a. On the importance of random error in the study of probability judgment. Part II: Applying the stochastic judgment model to detect systematic trends. *J. Behav. Decision Making* **10** 172–188.

Budescu, D. V., I. Erev, T. S. Wallsten, J. F. Yates, eds. 1997b. Introduction to special issue: Stochastic and cognitive models of confidence. *J. Behav. Decision Making* **10** 153–285.

Budescu, D. V., K. M. Kuhn, K. M. Kramer, T. Johnson. 2002. Modeling certainty equivalents for imprecise gambles. *Organ. Behav. Human Decision Processes* **88** 748–768.

Daniel, K. D., D. Hirshleifer, A. Subrahmanyam. 1998. Investor psychology and investor under- and overreactions. *J. Finance* **53** 1839–1886.

Daniel, K. D., D. Hirshleifer, A. Subrahmanyam. 2001. Mispricing, covariance risk, and the cross section of security returns. *J. Finance* **5** 921–965.

De Bondt, W. F. M., R. Thaler. 1995. Financial decision making in markets and firms: A behavioral perspective. R. A. Jarrow, V. Maksimovic, W. T. Ziemba, eds. *Finance, Handbooks in Operations Research and Management Science*, Vol. 9, Chap. 13. North-Holland, Amsterdam, The Netherlands, 385–410.

Du, N., D. V. Budescu. 2005. The effects of imprecise probabilities and outcomes in evaluating investment options. *Management Sci.* **51** 1791–1803.

Edwards, W. 1968. Conservatism in human information processing. B. Kleinmuntz, ed. *Formal Representation of Human Judgment*. Wiley, New York, 17–52.

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

Gigerenzer, G., U. Hoffrage, H. Kleinbolting. 1991. Probabilistic mental models: A Brunswikian theory of confidence. *Psych. Rev.* **98** 506–528.

Gilovich, T., D. Griffin, D. Kahneman. 2002. *Heuristics and Biases: The Psychology of Intuitive Judgment*. Cambridge University Press, Cambridge, UK.

Glaser, M., T. Langer, M. Weber. 2003. On the trend recognition and forecasting ability of professional traders. CEPR Discussion Paper 3904, Center for Economic Policy Research, London, UK.

Hoch, S. J. 1985. Counterfactual reasoning and accuracy in predicting personal events. *J. Experiment. Psych.: Learn., Memory, Cognition* **11** 719–731.

Jørgensen, M., K. H. Teigen, K. Moløkken. 2004. Better sure than safe? Overconfidence in judgment based software development effort prediction intervals. *J. Systems Software* **70** 79–93.

Juslin, P. 1993. An explanation of the hard-easy effect in studies of realism of confidence in one's general knowledge. *Eur. J. Cognitive Psych.* **5** 55–71.

Juslin, P. 1994. The overconfidence phenomenon as a consequence of informal experimenter guided selection of almanac items. *Organ. Behav. Human Decision Processes* **57** 226–246.

Juslin, P., P. Wennerholm, H. Olsson. 1999. Format-dependence in subjective probability calibration. *J. Experiment. Psych.: Learn., Memory, Cognition* **25** 1038–1052.

Juslin, P., A. Winman, H. Olsson. 2000. Naive empiricism and dogmatism in confidence research: A critical examination of the hard-easy effect. *Psych. Rev.* **10** 384–396.

Kahneman, D., P. Slovic, A. Tversky. 1982. *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press, Cambridge, UK.

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

Kirchler, E., B. Maciejovsky. 2002. Simultaneous over- and underconfidence: Evidence from experimental asset markets. *J. Risk Uncertainty* **25** 65–85.

Klayman, J. 1995. Varieties of confirmation bias. J. Busemeyer, R. Hastie, D. L. Medin, eds. *Decision Making from a Cognitive Perspective*. Academic Press, New York, 365–418.

Klayman, J., J. Soll, C. Gonzalez-Vallejo, S. Barlas. 1999. Overconfidence: It depends on how, what, and whom you ask. *Organ. Behav. Human Decision Processes* **79** 216–247.

Koriat, A., S. Lichtenstein, B. Fischhoff. 1980. Reasons for confidence. *J. Experiment. Psych.: Human Learn. Memory* **6** 107–118.

Kuhn, K. M., D. V. Budescu. 1996. The relative importance of probabilities, outcomes, and vagueness in hazard risk decisions. *Organ. Behav. Human Decision Processes* **68** 301–317.

Lawrence, M., P. Goodwin, M. O'Connor, D. Önkal. 2006. Judgmental forecasting: A review of progress over the last 25 years. *Internat. J. Forecasting* **22** 493–518.

Lichtenstein, S., B. Fischhoff. 1977. Do those who know more also know more about how much they know?: The calibration of probability judgments. *Organ. Behav. Human Performance* **20** 159–183.

Lichtenstein, S., B. Fischhoff, L. D. Phillips. 1982. Calibration of probabilities: The state of the art to 1980. D. Kahneman, P. Slovic, A. Tversky, eds. *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press, Cambridge, UK, 306–334.

Murphy, A. H. 1973. A new vector partition of the probability score. *J. Appl. Meteorology* **12** 595–600.

Rottenstreich, Y., A. Tversky. 1997. Unpacking, repacking, and anchoring: Advances in support theory. *Psych. Rev.* **104** 406–415.

Russo, J. E., P. J. H. Schoemaker. 1992. Managing overconfidence. *Sloan Management Rev.* **33** 7–17.

Seaver, D. A., D. V. Von Winterfeldt, W. Edwards. 1978. Eliciting subjective probability distributions on continuous variables. *Organ. Behav. Human Performance* **21** 352–379.

Soll, J. B. 1996. Determinants of overconfidence and miscalibration: The roles of random error and ecological structure. *Organ. Behav. Human Decision Processes* **65** 117–137.

Soll, J. B., J. Klayman. 2004. Overconfidence in interval estimates. *J. Experiment. Psych.: Learn., Memory, Cognition* **30** 299–314.

Teigen, K. H., M. Jørgensen. 2005. When 90% confidence intervals are 50% certain: On the credibility of credible intervals. *Appl. Cognitive Psych.* **19** 455–475.

Tversky, A., D. Kahneman. 1974. Judgments under uncertainty: Heuristics and biases. *Science* **185** 1124–1131.

Von Winterfeldt, D., W. Edwards. 1986. *Decision Analysis and Behavioral Research*. Cambridge University Press, Cambridge, UK.

Winman, A., P. Hansson, P. Juslin. 2004. Subjective probability intervals: How to reduce overconfidence by interval evaluation. *J. Experiment. Psych.: Learn., Memory Cognition* **30** 1167–1175.

Yaniv, I., D. P. Foster. 1995. Graininess of judgment under uncertainty: An accuracy- informativeness trade-off. *J. Experiment. Psych.: General* **124** 424–432.

Yaniv, I., D. P. Foster. 1997. Precision and accuracy of judgmental estimation. *J. Behavioral Decision Making* **10** 21–32.

Yates, J. F. 1982. External correspondence: Decomposition of the mean probability score. *Organ. Behav. Human Performance* **30** 132–156.

Yates, J. F. 1990. *Judgment and Decision Making*. Prentice Hall, Englewood Cliffs, NJ.