Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Unpacking the Future: A Nudge Toward Wider Subjective Confidence Intervals

Kriti Jain, Kanchan Mukherjee, J. Neil Bearden, Anil Gaba

To cite this article:
Kriti Jain, Kanchan Mukherjee, J. Neil Bearden, Anil Gaba (2013) Unpacking the Future: A Nudge Toward Wider Subjective Confidence Intervals. Management Science 59(9):1970-1987. https://doi.org/10.1287/mnsc.1120.1696

MANAGEMENT SCIENCE
Vol. 59, No. 9, September 2013, pp. 1970–1987
ISSN 0025-1909 (print) | ISSN 1526-5501 (online)

http://dx.doi.org/10.1287/mnsc.1120.1696

Unpacking the Future: A Nudge Toward Wider Subjective Confidence Intervals

Kriti Jain
Decision Sciences, INSEAD, Singapore 138676, kriti.jain@insead.edu

Kanchan Mukherjee
Indian Institute of Management Bangalore Bangalore 560076, India, kanchan.mukherjee@iimb.ernet.in

J. Neil Bearden, Anil Gaba
Decision Sciences, INSEAD, Singapore 138676 {neil.bearden@insead.edu, anil.gaba@insead.edu}

Subjective probabilistic judgments in forecasting are inevitable in many real-life domains. A common way to obtain such judgments is to assess fractiles or confidence intervals. However, these judgments tend to be systematically overconfident. Further, it has proved particularly difficult to debias such forecasts and improve the calibration. This paper proposes a simple process that systematically leads to wider confidence intervals, thus reducing overconfidence. With a series of experiments, including with professionals, we show that unpacking the distal future into intermediate more proximal futures systematically improves calibration. We refer to this phenomenon as the time unpacking effect, find it is robust to different elicitation formats, and address the possible reasons behind it. We further show that this results in better overall forecasting performance when improved calibration is traded off against less sharpness, and that substantive benefits can be obtained even from just one level of time unpacking.

Key words: confidence intervals; overconfidence; time unpacking
History: Received October 10, 2011; accepted October 5, 2012, by Teck Ho, decision analysis. Published online in *Articles in Advance* April 4, 2013.

1. Introduction

In many real-life situations, especially in the socioeconomic domains, prediction of future events necessarily involves subjective probabilistic judgments since well-defined data-generating processes simply do not exist. For example, experts are often called upon to forecast the price of oil or the level of Dow Jones Industrial Average at a future point in time, and they use their subjective judgments. For uncertain quantities that can take on a continuum of possible values, subjectively eliciting a complete probability distribution is an extremely difficult cognitive task not just for untrained individuals but also for those experienced in assessing probabilities. A common way to assess uncertainty in such cases is to obtain subjective assessments of *fractiles* or *confidence intervals*. Let X be an uncertain quantity that can take on a continuum of possible values. The kth $(0 \le k \le 1)$ subjective fractile is then the number X_k such that the subjective probability that the assessor assigns to X being below X_k is k, that is, $\mathrm{P}(X \le X_k) = k$. A subjective $100(1-\gamma)\%$, $0 \le \gamma \le 1$, confidence interval (CI) is a "low guess" and a "high guess" such that the assessor assigns probability $1-\gamma$ to the realized value of X being between the low guess and the high guess. The width of a CI is simply the interfractile range from two fractiles. For example, the width of a 90% CI for X is often represented as the interfractile range $(X_{0.05}, X_{0.95})$.

In the extensive literature on such subjective probabilistic judgments, since the 1970s and some even earlier, one of the most pervasive findings is that assessors display *overconfidence*: the assessed probability distributions tend to be too tight and hence are *miscalibrated* (see, e.g., Alpert and Raiffa 1982, Lichtenstein et al. 1982, Klayman et al. 1999, Griffen and Brenner 2004, Soll and Klayman 2004, Teigen and Jørgensen 2005). In other words, the subjective CIs tend to be too narrow, and the assessed extreme fractiles show a systematic bias toward underestimation of tail probabilities. For example, 90% subjective CIs for uncertain quantities are likely to capture much less than 90% of the actual realizations, often only 40% to 70% of the realizations. Many of the studies mentioned above used general knowledge questions as their stimuli. However, the overconfidence phenomenon has been replicated also in the context of forecasting real-life uncertain quantities. Deaves et al. (2010) report overconfidence among financial market practitioners in Germany who assessed 90% CIs for

the German market index DAX six months ahead. And, in most cases their CIs captured the realized value only between 40% and 70% of the time. Budescu and Du (2007) examine the quality of confidence judgments regarding future prices of financial assets and find overconfidence with 90% CIs, however not with 70% and 50% CIs. Russo and Schoemaker (1992) document overconfidence in CIs with business managers, where 90% CIs captured the true values between 42% and 62% of the time, while 50% CIs had included the true values about 20% of the time.

Does it really matter that the assessors are not well calibrated? Lichtenstein et al. (1982), for example, point out that if such assessments are made at several levels of an analysis, with each assessed distribution being too narrow, the errors instead of cancelling each other could compound, possibly leading to enormous expected losses. Barberis and Thaler (2003, p. 1063), in their survey of behavioral finance, underline overconfidence as one of the systematic biases of particular interest: "the confidence intervals people assign to their estimates of quantities—the level of the Dow in a year, say—are far too narrow." Also, they associate it with real economic consequences of financial losses and stock market volatility. It should not be difficult to imagine that overly narrow confidence intervals or underestimation of tail probabilities can have potentially catastrophic consequences, such as the most recent economic downturn (see, e.g., Johnson and Fowler 2011).

More troubling, it has proved particularly difficult to debias forecasts and improve the calibration of expressed subjective uncertainty. The most common mechanisms tried for reducing overconfidence have been feedback, training, and incentive schemes such as *scoring rules*, which have all yielded mixed results at best (in addition to Alpert and Raiffa 1982 and Lichtenstein et al. 1982 mentioned above, see, e.g., Hogarth 1975, Arkes et al. 1987, Koriat et al. 1980). The most common current prescription appears to be more or less that assessors should be careful about the overconfidence bias, although there has been some interesting recent work in this area. Soll and Klayman (2004) show that eliciting fractiles separately led to better calibration than asking directly for confidence intervals. Similarly, Winman et al. (2004) find strong evidence for format dependence. Participants in their studies gave interval estimates that were poorly calibrated (i.e., overconfident), yet the probabilities assigned to given intervals showed little bias. Haran et al. (2010) show that forcing forecasters to consider *all* possible outcomes for an event and then to assign probabilities to each significantly improves calibration relative to interval estimates. In this paper, we attempt to complement such efforts for improving calibration and thus reducing overconfidence in subjective confidence intervals.

We build upon intuition from earlier work such as *support theory* proposed by Tversky and Koehler (1994) and the stream of research on observed subadditivity of subjective probabilities (see, e.g., Ayton 1997, Bearden et al. 2007, Fox and Tversky 1998). This line of research shows that different descriptions of the same event lead to systematically different probabilistic judgments of the event. For instance, describing an event by its constituent pieces systematically leads to a higher subjective probability assigned to the event compared with when the event is not unpacked into its constituent pieces. So, for example, the sum of subjectively assessed probabilities for a person getting various types of cancer (such as liver cancer or lung cancer or pancreatic cancer, etc.) tends to be greater than the subjective probability for the same person getting any type of cancer when directly assessed (without breaking up the category of cancer into its constituent parts). Tversky and Koehler (1994) argued that dividing an event into its constituent parts tends to foster the accrual of support for the event and thereby increase the assessed subjective probability for that event.

Similarly, we hypothesized that unpacking a time horizon into its constituent pieces can lead to systematically different perceptions of time and can hence systematically influence judgments that involve that time horizon, such as subjective forecasts of quantities. More precisely, we conjectured that the perceived distance between now and some future point in time t is greater when one first explicitly considers some intermediate points in time t_1 and t_2, where $t_1 < t_2 < t$, than when one considers t directly. For example, simply making it salient that between now and one year ahead there is three months ahead and six months ahead would make one year ahead seem further out. This then should increase the uncertainty in an assessor's forecast for an event or a quantity one year ahead. In other words, *unpacking time* in this manner should lead to, for example, wider subjective confidence intervals. This is indeed what we show.

In an extensive series of experiments, with MBA students and with financial analysts in their domain of expertise at a large international brokerage and investment group and at a large global asset management firm, we find that unpacking time into its constituent pieces tends to systematically increase the uncertainty assessors have in their forecasts, which can then potentially mitigate the overconfidence in those forecasts. Say we wish to elicit an analyst's 90% CI for the value of Dow Jones Industrial Average three months from today. The typical approach is to simply ask for the CI directly for the desired time, which in this case is a 90% CI three months hence. Our research consistently shows that asking assessors first to give 90% CIs for one month out and two months out before

doing so for three months out can substantially widen the intervals at three months and thereby potentially improve calibration. We refer to this phenomenon as the *time unpacking effect*. We show the unpacking effect in several experiments using different events and elicitation procedures.

In one of our experiments, we asked assessors for three-month forecasts for several real-world uncertain quantities under two conditions. Assessors in the *packed condition* gave CIs only for three months out, and those in the *unpacked condition* gave CIs for one and two months out before giving CIs for three months out. The mean width of the CIs was significantly greater in the unpacked condition. In a related experiment, we observe similar time unpacking effects when subjects were asked to estimate the probabilities of future events (e.g., the probability that the Dow will increase/decrease by 15% or more at least once in the next t months). The estimates indicate greater perceived volatility when some intermediate estimations are made on an unpacked timeline compared with direct estimates in a packed timeline. We also report experiments designed to address possible cognitive mechanisms that lead to the time unpacking effect. We think that one possibility for this effect is due to changes in the perception of time that can occur because of the psychophysics of time perception (e.g., Stevens 1975). Specifically, the time perception is nonlinear and compressed (the distance of 10 years ahead is perceived to be less than twice the distance of 5 years ahead, for example) and unpacking time helps decompress it. Another possibility is that unpacking time fosters relative comparisons that influence perceived time distance—for example, 10 years ahead seems further out when explicitly compared with 3 years and 5 years than when 10 years ahead is considered in isolation. Of course, both these mechanisms may play a role, rather than one or the other. We hope that by understanding the underlying mechanisms that drive the time unpacking effect we can develop better aids to improve forecasting.

In all the experiments mentioned above, we show that the time unpacking effect leads to assessments of greater uncertainty (wider confidence intervals or higher tail probabilities) that should presumably improve calibration. However, to actually measure the calibration in those experiments, one would need a large number of repeated trials, which is an onerous task in the context of real-life assessments that we did. So, we also conducted an experiment where the data-generating process was well defined and known and hence where it was possible to get precise measures of calibration, and we then also examined the impact of different levels of time unpacking (dividing the target time into different number of intermediate constituent pieces) on CI widths and consequent calibration. We find that although assessors remain overconfident even with time unpacking, the degree of overconfidence decreases. And, the most substantial and significant improvement comes with just one level of unpacking, with further marginal improvements from additional levels of time unpacking happening at a decreasing rate. Such improvements in calibration do not seem to come at a cost of lower overall forecasting performance that accounts also for sharpness. We evaluate the overall forecasting performance with a scoring rule that takes into account a trade-off between improved calibration and less *sharpness* (i.e., increased CI widths). The scores improve with time unpacking and, consistent with earlier results, the most benefit comes from just one level of unpacking with the marginal rate of improvement decreasing with additional levels of unpacking. To sum up, we observe that the wider CIs from time unpacking lead to improved calibration without creating underconfidence, and the improved calibration when traded off against worsening sharpness still results in a better overall forecasting performance. And, just one level of time unpacking leads to the most substantial and significant benefit in terms of improved forecasting performance.

In §2, we specify the class of stochastic processes that we consider and briefly describe a model for biased predictions. In §3, we discuss Experiments 1 and 2 that show the basic time unpacking effect through CI elicitations with MBAs and with real-life professionals in their field of expertise. In §4, we explore with Experiment 3 the possible causal mechanisms of the unpacking effect. In §5, through Experiments 4a and 4b, we show that time unpacking is robust to different elicitation formats. In §6, we describe Experiment 5 where we compute calibration of assessed confidence intervals under the packed and unpacked conditions, and consider the impact of different levels of time unpacking on CI widths as well as on the resulting calibration and the overall forecasting performance that accounts for more than just calibration. In §7, we conclude with a summary and discussion.

2. A Model for Overconfidence in Subjective Confidence Intervals

Let $\{X(t), 0 \le t < \infty\}$ be a collection of random variables, that is, a stochastic process with $X(t)$ as the state of the process at time t. We consider the class of stochastic processes where the shortest prediction interval for $X(t)$ is increasing in t. In formal terms, assuming without loss of generality that the process is at $t = 0$, the shortest prediction interval for $X(t)$, $t > 0$, given a preassigned probability, say $1 - \gamma$, is given by $[a(t), b(t)]$ such that

$$1. \quad \text{Prob}\{a(t) \le X(t) \le b(t)\} = 1 - \gamma, \quad \text{and} \tag{1}$$

$$2. \quad b(t) - a(t) \text{ is minimized.} \tag{2}$$

Then we consider the class of stochastic processes in which the width of this interval, $b(t)-a(t)$, is increasing in t for all γ. This is a way to specify that the uncertainty about$X(t)$ is increasing in t.

This is a broad class of stochastic processes, and includes processes that have been widely applied, for example, to simulate behavior of stock prices, currencies, futures, interest rates, and other financial quantities (see, e.g., Hull 2008 and McDonald 2005). For instance, one such stochastic process is the commonly used Brownian motion process in which, under the condition that $X(0)=0$, at $t=0$, $X(t)$ for $t>0$ is normally distributed with mean 0 and variance $\sigma^2 t$, with σ as a fixed parameter (see, e.g., Ross 1983). In this case, the shortest prediction interval at $t=0$ for $X(t)$, $t>0$, given a preassigned probability $1-\gamma$ is defined by

$$\text{Prob}\{-z_{\gamma/2}\sigma\sqrt{t} \le X(t) \le z_{\gamma/2}\sigma\sqrt{t}\} = 1-\gamma, \quad (3)$$

where $z_{\gamma/2}$ is the $100(\gamma/2)$th percentile of the standard normal distribution. Then the width of this interval, which is $2z_{\gamma/2}\sigma\sqrt{t}$, is increasing in t.

Now, imagine an assessor at $t=0$ providing a subjective confidence interval (equivalent to shortest prediction interval) for $X(t)$, $t>0$. A well-established finding from the psychophysics of time perception is that time is (cognitively) represented nonlinearly (see, e.g., Stevens 1975, Eisler 1976). More precisely, representations of time tend to be compressed such that the perception of time t into the future is $R(t)<t$. Thus, $R(t)<t$ makes t *seem* closer than it is. For example, the distance of 10 years ahead is perceived to be less than twice the distance of 5 years ahead. One commonly used representation of $R(t)$ is $R(t)=t^{\alpha}$, where $0<\alpha<1$.

Then, within the class of stochastic processes that we consider, subjective assessments of confidence intervals are biased. In other words, an assessor providing subjective confidence intervals for $X(t)$, $t>0$, perceives the time period t to be less than what it truly is, as $R(t)<t$. And, hence, the assessor perceiving $X(t)$ to be $X(R(t))$ underestimates the uncertainty (i.e., the width of the confidence interval) associated with $X(t)$. So, for a preassigned probability the assessor judges a confidence interval for $X(t)$ to be tighter than it should be, and equivalently, for a preassigned interval of $X(t)$ judges a higher likelihood than it should be. In sum, the *compressed* perception of time leads the assessor to be overconfident.

We do not think or claim that such a compressed perception of time is the only source of overconfidence. Nevertheless, to the extent that this does contribute to overconfidence, it seems worthwhile to try and deal with it. Further, our approach is limited to forecasting problems that involve a time dimension and where the uncertainty about the quantity of interest is increasing with time. So, although the process need not be stationary, the shortest prediction interval must be increasing in time, which might not be the case in contexts such as sales with a strong seasonality factor or other similar confounding variables.

3. The Time Unpacking Effect in Subjective Confidence Intervals

3.1. Experiment 1: Method

Experiment 1 was designed to check for the existence of the time unpacking effect. The participants were 128 MBA students at INSEAD with an average age of about 29 years who were randomly assigned into the *packed* and *unpacked* conditions. In both conditions, the participants were given the last available closing level of three financial indicators: the Dow Jones Industrial Average (Dow), the price of Brent crude oil in U.S. dollars (Oil), and the Google stock price (Google), and were asked for their 90% CIs at certain points in the future. In the packed condition, participants made predictions for the three quantities three months ahead; in the unpacked condition, participants made predictions for the same quantities one month, two months, and three months ahead, in that order. Our hypothesis was that the average interval width for three months would be wider in the unpacked condition compared with the packed condition.

The question in the unpacked condition is provided below for illustration:

> For the three quantities below, you are given the last closing price available. For three points in the future, one month, two months, and three months, we would like you to estimate your 90% confidence interval (i.e., an upper bound such that there is 5% probability that the quantity will close *above* that value and a lower bound such that there is a 5% probability that the quantity will close *below* that value; in other words, the interval you state should be such that you are 90% confident that the actual value will lie within that interval).

3.2. Experiment 1: Results

The critical comparison between the packed and unpacked conditions is the width of the 90% CIs at three months for the three quantities, where a CI width is defined as the difference between the upper bound and the lower bound. Table 1 shows for each of the three quantities the means and the standard errors of the means for the lower bounds, upper bounds, and the widths of the CIs under the unpacked and packed conditions, along with the percentage increase in the mean unpacked CI widths relative to the mean packed CI widths. Consistent with our hypothesis, for each of the three quantities, the mean CI width for the unpacked condition

Table 1 **90% CIs for Three Months Ahead Under the Packed and Unpacked Conditions: Means and Standard Errors of the Means (SEM) for Lower Bounds (LB), Upper Bounds (UB), and CI Widths, Along with Percentage Increase in Mean CI Widths in the Unpacked Condition vs. the Packed Condition (Experiment 1)**

	Mean (SEM)						
	Packed ($N = 68$)			Unpacked ($N = 60$)			% increase in mean CI width, unpacked vs. packed
	LB	UB	CI width	LB	UB	CI width	
Dow	8,962 (160)	11,594 (141)	2,632 (319)	8,521 (226)	12,017 (235)	3,497 (401)	33
Oil	60 (1)	95 (2)	35 (3)	57 (2)	104 (3)	47 (5)	34*
Google	473 (10)	650 (9)	177 (18)	452 (18)	688 (15)	237 (31)	33

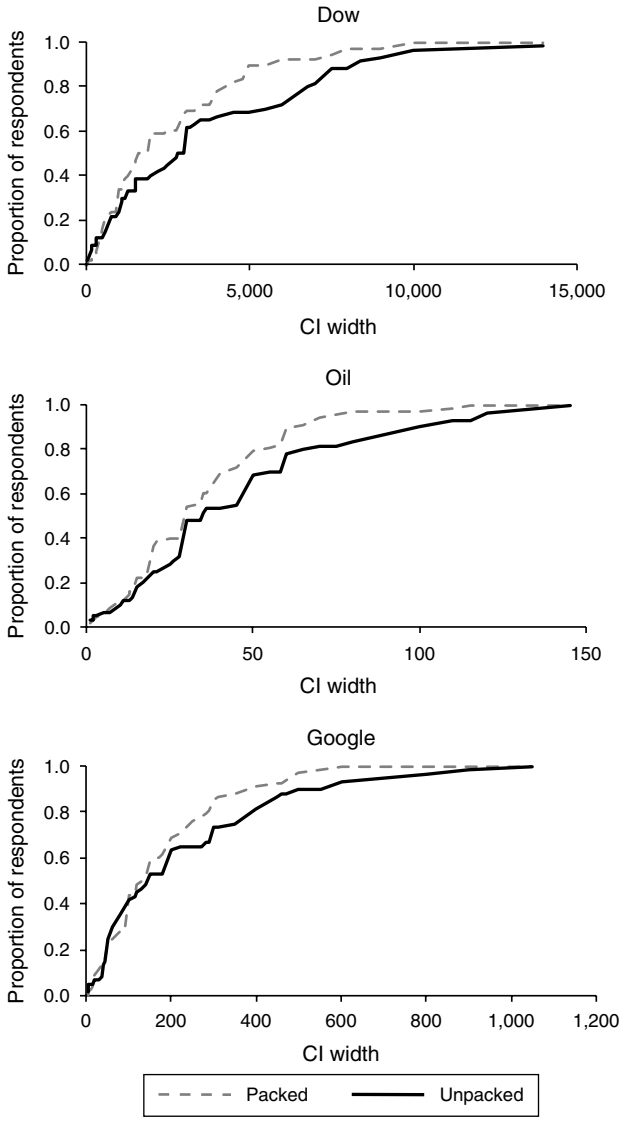
*$p < 0.05$ (two-tailed).

was significantly greater than that for the packed condition. This increase in mean CI width was about 34% uniformly across the three quantities. In the case of Dow, $t(109.4) = 1.75$, p(two-tailed) $= 0.08$; for Oil, $t(101.4) = 2.16$, p(two-tailed) $= 0.03$; and for Google, $t(94.72) = 1.68$, p(two-tailed) $= 0.09$.[1] Note also that, for each of the three quantities, the mean lower bound (mean upper bound) in the unpacked condition is less (greater) than that in the packed condition. In this sense of on average, the packed CI is a subset of the unpacked CI. In other words, on average, an unpacked CI subsumes the uncertainty in the corresponding packed CI and then adds some more.

Figure 1 shows the cumulative distributions for the CI widths in the packed and unpacked conditions for the three quantities. For each of the three quantities, the unpacked distribution is almost always below the packed distribution. For example, in the case of DOW, roughly 70% of the CI widths in the unpacked condition were 5,000 or less, whereas about 90% of the CI widths in the packed condition were 5,000 or less. In other words, roughly 10% of the packed CI widths were greater than 5,000, whereas close to 30% of the unpacked CI widths were greater than 5,000. This seems to be almost always the case for any given CI width for each of the three quantities.

Assuming, in line with the extensive literature mentioned earlier, that calibrations of 90% CIs are well under 90%, the results in Table 1 and Figure 1 are suggestive that the unpacked intervals will likely lead to better calibration than the packed intervals. To actually measure the calibration of these intervals, given that the data generating processes for the forecasted real-life quantities are impossible to define, the analysis would require a large number of repeated trials over a long period of time. Later, in §6, we do show some calibration results in a setting where the data generating process is known and well defined. Still, to get some flavor here on potential calibration of the packed and unpacked intervals, we used past empirical data on the forecasted quantities to evaluate these intervals. More specifically, going back about five years from the day of the experiment (from January 1, 2006, to December 31, 2010), for each day,

Figure 1 **Cumulative Distributions for CI Widths in the Packed (F_P) and Unpacked (F_U) Conditions (Experiment 1)**



[1] One-sided t-tests yielded statistical significance at $p < 0.05$ in all the cases. However, as in the rest of the paper, we report the more conservative two-sided p-values.

we looked at the percentage change in each quantity (Dow, Oil, and Google) three months ahead. This gave us for each quantity a distribution for the percentage change three months ahead. Then, for a given assessed 90% CI in Experiment 1, we computed the implied 90% CI for percentage change in the underlying quantity. For example, if a subject stated a 90% CI for Dow three months ahead to be (9,500, 10,500), with the Dow being 10,000 (known to the subject) on the day of the assessment, the implied 90% CI for percentage change in Dow three months ahead was taken to be (−5%, 5%). Thereafter, we computed the proportion of the observations in the past five-year empirical distribution for the quantity that were captured by the implied 90% CI for that quantity. We refer to this capture rate as the five-year *trailing calibration*. The average five-year trailing calibration of packed versus unpacked intervals was 62% versus 63% for Dow, 64% versus 68% for Oil, and 47% versus 52% for Google. Measured this way, calibration in the unpacked condition vis-à-vis the packed condition increased by 1%, 6%, and 11% for Dow, Oil and Google, respectively. We replicated this also with 10-year trailing calibrations and obtained very similar results.

To sum up, these results are further indicative that overconfidence is a persistent issue in judgments of uncertainty and that unpacking time can play a role in reducing the degree of overconfidence. The improvements in the trailing calibration through unpacking time are small, but the results are consistent across quantities and could entail large economic consequences. It is a very simple procedure that can be useful as a robust complement to other efforts in reducing overconfidence. In the next experiment, we replicated this effect with professionals in the finance industry.

3.3. Experiment 2: Method

In this experiment we went to the field to explore whether professionals with substantive expertise in finance are also susceptible to the time unpacking effect. Seventy-three analysts at a major international brokerage and investment group, CLSA, participated in the experiment. CLSA is headquartered in Hong Kong, and has more than 1,500 professionals located in 15 major Asian cities and in other major financial centers such as London, New York, and Sydney (for more information, see http://www.clsa.com). The median age of the participants, all with university or advanced degrees, was 36 years, and the median years of service with CLSA (not including experience at other financial houses) was five years. The analysts in our sample were based in New York, Tokyo, and Hong Kong. They participated in an online survey where they made forecasts for some of the financial quantities that were of most interest to and were continuously tracked by them. The financial quantities in our study were the price of Oil Brent EUCRBRDT Index (Oil), the price of Gold US$/oz GOLDS Commodity (Gold), the Dow Jones Industrial Average Index (Dow), the Nikkei NKY Index (Nikkei), the Hang Seng HIS Index (Hang Seng), and the MSCI Asia Free × Japan MXASJ Index (MSCI).

As before, time unpacking was manipulated between subjects with two conditions, packed and unpacked. The procedure was the same as in Experiment 1. Participants in the packed condition were asked for 90% CI (lower and upper bounds) three months ahead and those in the unpacked condition were asked for projections one month, two months, and three months ahead, in that order. All analysts made their predictions within a 24-hour time frame on April 1, 2010.

3.4. Experiment 2: Results

There was a systematic tendency for analysts in the unpacked condition to give wider intervals than in the packed condition across all six quantities. Table 2 shows the means and the standard errors of the means for lower bounds, upper bounds, and widths of the 90% CIs for three months ahead for all the six quantities in the packed and unpacked conditions, along with the percentage increase in the mean CI widths in the unpacked condition relative to the packed condition. The increase in the mean CI width in the unpacked condition relative to the packed condition ranged from 20% to 44% with an average increase of 31% across the six quantities. None of the paired comparisons in Table 2 are statistically significant, but the consistently larger CI widths in the unpacked condition across all the six quantities suggest that the time unpacking effect persists also with professionals in the field predicting quantities they work with in their profession.[2] Also, the average lower bounds (upper bounds) are smaller (larger) in the unpacked condition than in the packed condition, except for one quantity (MSCI) where the average lower bounds are the same under the two conditions. In this sense, as in Experiment 1, on average, an unpacked CI for an uncertain future quantity envelops the corresponding packed CI.

Hence, the findings in the MBA sample suggesting assessments of greater uncertainty in an unpacked CI relative to a packed CI, and thereby potentially better calibration of unpacked CIs vis-à-vis packed CIs, translate to this sample from the professionals as well. As with the results of Experiment 1, we also looked at the 5-year and 10-year trailing calibrations of the packed and unpacked intervals. The improvements in the trailing calibration in the unpacked intervals

[2] Note that the probability of observing greater CI widths in the unpacked condition vis-à-vis the packed condition across all six quantities due to just a random occurrence is $0.5^6 = 0.016$.

Table 2 **90% CIs for Three Months Ahead Under the Packed and Unpacked Conditions: Means and Standard Errors of the Means (SEM) for Lower Bounds (LB), Upper Bounds (UB), and CI Widths, Along with Percentage Increase in Mean CI Widths in the Unpacked Condition vs. the Packed Condition (Experiment 2)**

	Mean (SEM)						
	Packed ($N = 42$)			Unpacked ($N = 31$)			% increase in mean CI width
	LB	UB	CI width	LB	UB	CI width	
Oil	74 (1)	93 (2)	19 (2)	71 (3)	97 (3)	26 (5)	37
Gold	1,038 (21)	1,241 (26)	203 (25)	1,018 (29)	1,264 (32)	246 (56)	21
DJIA	9,965 (133)	11,647 (115)	1,682 (225)	9,806 (285)	12,069 (268)	2,263 (509)	35
Nikkei	10,391 (123)	12,238 (144)	1,847 (234)	10,162 (287)	12,814 (291)	2,652 (530)	44
Hang Seng	19,690 (280)	22,891 (157)	3,201 (371)	19,476 (499)	23,317 (486)	3,840 (737)	20
MSCI	445 (10)	542 (11)	97 (12)	445 (17)	569 (18)	125 (25)	28

as compared with the packed intervals were small, however. For example, the average five-year trailing calibration across all six quantities was 47% for the unpacked intervals compared with 45% for the packed intervals, the best improvement being for Oil (46% for unpacked versus 41% for packed, a 12% increase) and no improvement in the case of Gold. The 10-year trailing calibrations were very similar. But, as in Experiment 1, the results are consistent across the various forecasted quantities.

We investigate the possible causes of the time unpacking effect in the next section.

4. Why Does Time Unpacking Lead to Wider Confidence Intervals?

In §2, we discuss how representations of time tend to be compressed such that the perception of time t into the future is $R(t) < t$. Thus, $R(t)$ makes t *seem* closer. By this account, making explicit some of t's constituent components, $t_1, t_2, \ldots, t_m$, where $t = t_1 + (t_2 - t_1) + \cdots + (t_m - t_{m-1})$, can effectively decompress a forecaster's perception of time to $R(t_1) + R(t_2 - t_1) + \cdots + R(t_m - t_{m-1}) > R(t)$, thereby causing the unpacked time horizon to seem longer than its packed counterpart. We refer to this potential explanation as the *representational account of time*.

An alternative explanation is that unpacking time fosters the use of relative comparisons—or contrasts—which then drive the observed time unpacking effects. Helson (1964) argued—with the support of substantial persuasive data—that prior experience with a stimulus provides a reference point against which subsequent stimuli are judged (for related findings, see, e.g., Beebe-Center 1929, Manstead et al. 1983). For instance, he found that judgments for the heaviness of objects depended on the order in which the objects were judged. An object of a fixed weight felt heavier after a light object than after a heavy object. So, just as a 5 kg object will be perceived to be lighter when judged after a 10 kg object than after a 2 kg object, a time t might be perceived as further into the future after considering $t_{\text{near}} < t$ than after $t_{\text{far}} > t$. Such an effect from relative comparisons is now commonly referred to as a *contrast effect*.

The representational account of time and the contrast effect can lead to distinctly different predictions. Suppose, for instance, that a judge forecasts an event for time periods t_1, t_2, and t_3, a second judge forecasts the same event for time period t_3, and a third judge does so for time periods t_3, t_4, and t_5, where $t_1 < t_2 < t_3 < t_4 < t_5$. By the representational account of time, t_3 should be perceived similarly by the second and third judges, and relative to them, further out in the future by the first judge. However, by the contrast effect, the first judge relative to the second judge should perceive t_3 to be further out in time (since t_3 is being compared with shorter time periods t_1 and t_2 by the first judge rather than being considered by itself as by the second judge), but the third judge relative to the second judge should perceive t_3 to be nearer (since t_3 is being compared with longer time periods t_4 and t_5 by the third judge).

We report results from an experiment that allow us to determine the relative descriptive adequacy of these two potential explanations.

4.1. Experiment 3: Method

The participants were 216 MBA students at INSEAD. These participants were distinct from the participant pool in Experiment 1. They were divided randomly into three conditions, *packed, unpacked*, and *extended*, in a between subjects design. The packed and unpacked conditions, and the quantities to be forecasted (Dow, Oil, Google), were identical to those in Experiment 1: in the packed condition, subjects gave 90% CIs for three months ahead; in the unpacked condition, subjects gave 90% CIs for one month, two months, and three months ahead, in that order. In the extended condition, participants were asked to give 90% CIs for the same three quantities, but for three months, four months, and five months ahead, in that order. By the contrast effect, if the presence of the one month and two months stimuli in the unpacked condition causes three months to seem longer than in the packed case, then the presence of the four and five months stimuli in the extended

condition should make three months seem shorter than in the packed case. On the other hand, by the representational account of time, three months should be perceived identically in both the packed and extended conditions, whereas it should be perceived longer in the unpacked condition, as discussed above. In other words, with $W(\cdot)$ representing the CI width, W(unpacked) > W(packed) > W(extended) provides support for the contrast effect, whereas W(unpacked) > W(packed) = W(extended) lends support to the representational account of time.

4.2. Experiment 3: Results

Table 3 shows the summary results for the 90% CIs three months ahead for the three quantities: means and standard errors of the means for lower bounds, upper bounds, and CI widths, along with the percentage increases in CI widths in pairwise comparisons of the three conditions (packed, unpacked, and extended). We conducted three separate analysis of variance (ANOVA) using CI widths of the three quantities (Dow, Oil and Google) as dependent measures and the three experimental conditions (packed, unpacked, and extended) as a between-subjects factor. Across all three dependent variables, ANOVAs demonstrated a significant main effect of the conditions. For each of the three variables, the mean CI width for the unpacked condition was greater than the mean CI width for the packed and extended conditions, and the mean CI width for the packed condition was greater than that for the extended condition. For Dow, the mean CI width in the unpacked condition was 27% greater than that in the packed condition ($t(128.3) = 1.49$, p(two-tailed) = 0.14) and 81% greater than that in the extended condition ($t(120.1) = 3.15$, p(two-tailed) = 0.002); whereas the mean CI width in the packed condition was 43% greater than that in the extended condition ($t(142) = 1.87$, p(two-tailed) = 0.06). In the case of Oil, the mean CI width in the unpacked condition was 44% larger than that in the packed condition ($t(127.1) = 2.17$, p(two-tailed) = 0.03) and 86% larger than that in the extended condition ($t(95.2) = 3.74$, p(two-tailed) < 0.001); the mean CI width in the packed condition was 29% larger than that in the extended condition ($t(99) = 1.6$, p(two-tailed) = 0.12). Finally, for Google, the mean CI width in the unpacked condition was 42% greater than that in the packed condition ($t(126.3) = 2.50$, p(two-tailed) = 0.01) and 83% greater than that in the extended condition ($t(106.4) = 4.22$, p(two-tailed) < 0.001); the mean CI width in the packed condition was 29% greater than in that in the extended conditions ($t(115.3) = 1.8$, p(two-tailed) = 0.06).[3] Note also in

[3] One-sided t-tests yield statistical significance at $p < 0.05$ in eight of the nine pairwise contrasts, with the ninth being directionally consistent.

Table 3 Summary Statistics for 90% CIs Three Months Ahead Under the Packed, Unpacked, and Extended Conditions: Means and Standard Errors of the Means (SEM) for Lower Bounds (LB), Upper Bounds (UB), and CI Widths, Along with Percentage Increases in Mean CI Widths in Pairwise Comparisons of the Three Conditions (Experiment 3)

	Mean (SEM)									% increase in mean CI width		
	Packed ($N = 63$)			Unpacked ($N = 72$)			Extended ($N = 81$)					
	LB	UB	CI width	LB	UB	CI width	LB	UB	CI width	Unpacked vs. packed	Unpacked vs. extended	Packed vs. extended
Dow	9,158 (94)	11,232 (119)	2,074 (209)	8,971 (86)	11,595 (182)	2,623 (312)	9,349 (93)	10,891 (91)	1,541 (180)	27	81**	43
Oil	63 (11)	90 (1)	27 (3)	60 (9)	98 (2)	39 (3)	64 (10)	85 (1)	21 (2)	44*	86**	29
Google	460 (18)	630 (23)	170 (16)	434 (19)	676 (31)	242 (25)	476 (14)	608 (22)	132 (12)	42**	83**	29

*$p < 0.05$; **$p < 0.01$ (two-tailed).

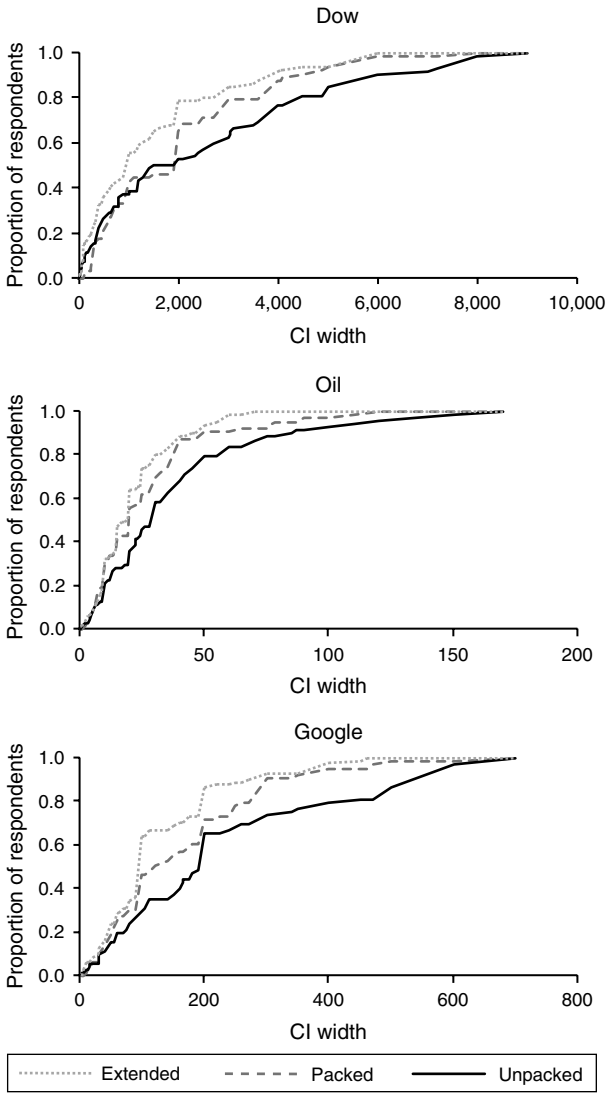
Table 3 that, for each of the three quantities, on average, LB (unpacked) < LB (packed) < LB (extended), and UB (unpacked) > UB (packed) > UB (extended). In this sense, on average, the extended CI is a subset of the packed CI, which in turn is a subset of the unpacked CI. The results in Table 3 are further validated in Figure 2, which shows the cumulative distributions of the CI widths under the packed, unpacked, and extended conditions for Dow, Google, and Oil.

Experiment 3, in addition to replicating the unpacking effect with the earlier experiments, lends support to the contrast effect of time. That is, comparing a time period of interest t with shorter and intermediate time periods makes t seem further out, leading to assessments of greater uncertainty associated with events at t.

How does this translate into levels of trailing calibration? We would expect these to be better as we go from the extended to the packed to the unpacked conditions, in line with the higher average CI widths. This is exactly what we see in Table 4, which shows the average trailing calibrations for each forecasted quantity under the packed, unpacked, and extended conditions. For each of the three forecasted quantities, the average trailing calibrations are the lowest under the extended condition and the highest under the unpacked condition, with the differences being fairly substantive. Once again, this is indicative of the persistent overconfidence, with time unpacking improving calibration with respect to packed time, and the extended time format leading to worse calibration relative to packed time.[4]

[4] Both the contrast effect and the representational account of time propose that the unpacking process alters the perception of time, which then translates into different judgments of uncertainty. As one reviewer pointed out, another possible mechanism could simply be that an assessor has some implicit estimate of uncertainty that she then adjusts for comparative assessments across different time periods. For example, the elicited judgment for the first time period could serve as an anchor that is then adjusted to account for greater uncertainty for greater time periods. If this kind of conscious comparative judgment is indeed the case, then asking assessors to make predictions for unrelated events for intermediate (in the unpacked condition) or extended (in the extended condition) time periods should make no difference to their assessments of uncertainty for the focal event at the time period of interest. We ran an experiment to test this comparative judgment account. Participants were divided randomly into three conditions, *packed, primed unpacked*, and *primed extended*, in a between-subjects design. Participants in the packed condition were asked to predict the number of emails they would receive in the next six months. Participants in the primed unpacked condition were asked to predict the number of emails they would receive in the next one, three, and six months, in that order. And, in the primed extended condition, participants made the same predictions for 6, 9, and 12 months, in that order. Next, in all conditions, participants assessed 90% CIs for Dow six months ahead. Finally, participants were asked to also give their subjective assessment of how far they felt six months was from now, using a 0 (very near) to 10 (very far) scale. Two separate ANOVAs using CI width of Dow and perception of time as the dependent measure and the three experimental conditions (packed, primed unpacked, and primed extended) as a between-subjects factor demonstrated a significant main effect of the conditions, with the dependent measures being the greatest (lowest) for the primed unpacked (primed extended) condition. These findings provide further support to our conjecture that unpacking alters the perception of time.

Figure 2 **Cumulative Distributions for CI Widths in the Packed (F_P), Unpacked (F_U), and Extended (F_E) Conditions (Experiment 3)**



5. Time Unpacking Effect with Elicitation Formats Other Than Confidence Intervals

Experiments 1, 2, and 3 investigated the existence of the time unpacking effect and possible underlying psychological mechanisms that lead to this effect. In this section, with Experiments 4a and 4b, we demonstrate that this effect is robust to alternative elicitation formats.

Table 4 Average Five-Year (2006–2010) Trailing Calibrations Under the Packed, Unpacked, and Extended Conditions (Experiment 3)

	Mean (%)			% increase in mean trailing calibration		
	Packed ($N = 63$)	Unpacked ($N = 72$)	Extended ($N = 81$)	Unpacked vs. packed	Unpacked vs. extended	Packed vs. extended
Dow	47	53	36	11	48	33
Oil	47	57	40	22	41	16
Google	43	54	35	24	52	23
Overall	46	54	37	17	46	24

5.1. Experiment 4a: Method

The participant pool consisted of 131 MBA students at INSEAD randomly divided into the packed, unpacked, and extended conditions, as in Experiment 3. Participants were asked to judge the likelihood of three pairs of financial events related to the three quantities that were used in Experiments 1 and 3 (Dow, Oil, and Google) for different time periods into the future: Dow being below and above its current level by 10% or more at least once, Oil being below and above its current price by 15% or more at least once, and Google stock price being below and above its current price by 20% or more at least once. For illustration, one of the three pairs of questions read as follows:

> On January 20, 2010, the price of *crude oil* (Brent) was *\$78.01* per barrel. What is the probability that the price will be *below* its current level by 15% or more *at least once* in the next X months? Next, what is the probability that the price will be *above* its current level by 15% or more *at least* once in the next X months?

As in Experiment 3, $X =$ one month, two months, and three months (in that order) in the unpacked condition; $X =$ three months in the packed condition; and $X =$ three months, four months, and five months (in that order) in the extended condition. Note that the questions were designed so that normatively the assessed subjective probability should increase, or at least not decrease, with time.

Typically, in judging probabilities for a given event, overconfidence is indicated by assessed probabilities being too near certainty (i.e., probabilities near 0 or 1) and underconfidence is reflected by assessed probabilities being nearer to 0.5 than they should be (see, e.g., Morgan et. al 1990, p. 110). However, our focus here is on tail probabilities of a distribution and hence our view of overconfidence is somewhat different. We hypothesized that, consistent with the time unpacking effect, respondents in the unpacked condition would display greater appreciation of uncertainty and hence assign greater probabilities to the tails of a distribution (i.e., greater probabilities to both the questions for each quantity stated above) compared with the respondents in the packed condition who in turn will assign greater probabilities than respondents in the extended condition.

5.2. Experiment 4a: Results

The mean probability judgments across the three conditions and for the two events for each of the three quantities (Oil, Dow, Google) are shown in Figure 3. Across the three quantities, the mean probabilities for each of the two tail events are the highest in the unpacked condition, next highest in the packed condition, and the lowest in the extended condition. Figure 4 shows for each of the six events the cumulative distributions of the assessed probabilities under the three conditions (packed, unpacked, and extended). We also ran separate ANOVA for each of the six target

Figure 3 Mean Probabilities in the Packed, Unpacked, and Extended Conditions (Error Bars Represent ±1 Standard Error of Means) (Experiment 4a)

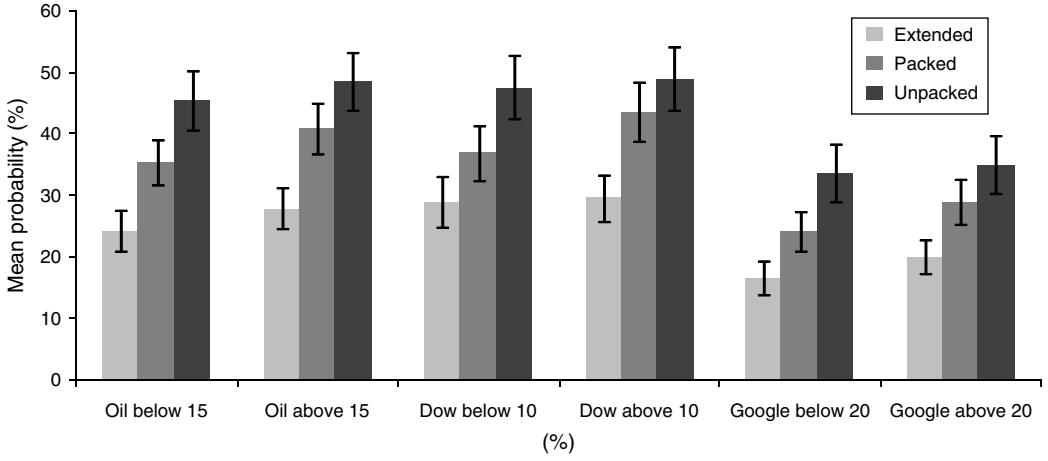
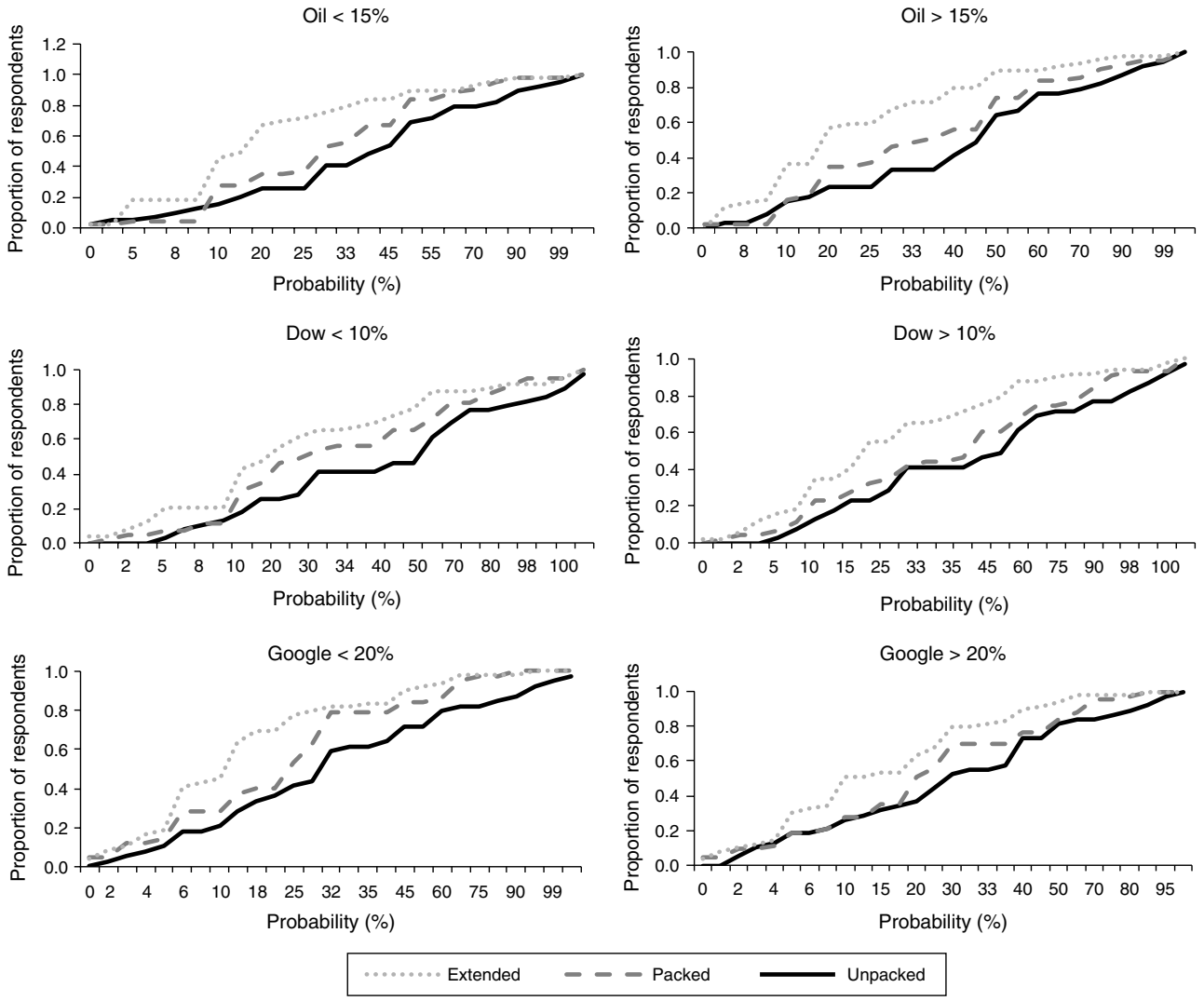


Figure 4 **Cumulative Distributions for Probabilities in the Packed (F_P), Unpacked (F_U), and Extended (F_E) Conditions (Experiment 4a)**



questions with the three experimental conditions as a between-subjects factor. ANOVAs revealed a significant main effect for the conditions; for reasons of space and given the consistent flavor of the data in Figures 3 and 4, we do not report the statistics here.

The data here suggests that our hypothesis on the time unpacking effect in the assessed CIs extends to elicitations of tail probabilities as well. We replicate this effect in yet another elicitation format for probabilistic judgments in the next experiment.

5.3. Experiment 4b: Method

In this experiment, instead of eliciting CIs or tail probabilities associated with percentage increase or decrease in uncertain quantities, we asked for the subjective probability associated with a specific interval of an uncertain quantity. We asked for the subjective probability that the price of Gold (US$/oz, GOLDS Commodity) will be between $1,050 and $1,350, 12 months ahead in the packed condition; and 3 months, 6 months, and 12 months ahead, in that order, in the unpacked condition.

Our hypothesis was that participants in the unpacked condition compared with those in the packed condition would assign *lower* subjective probabilities that Gold price would lie within the defined interval 12 months ahead, indicating fatter tail probabilities. The participant pool consisted of 62 analysts at a major international asset management firm (a client of the brokerage and investment group, CLSA) with an average age of 33 years, 7.7 years of experience on average in the finance industry, and representing 20 different nationalities in the Tokyo, London, and Hong Kong offices. They were randomly divided into the packed and unpacked conditions in a controlled online survey.

5.4. Experiment 4b: Results

As hypothesized, the mean probability judgment in the unpacked condition ($M = 32.8\%$, $SD = 20.0\%$,

$N = 29$) was significantly smaller than that in the packed condition ($M = 45.9\%$, SD $= 25.9\%$, $N = 33$), $t = 2.24$, p(two-tailed) $= 0.03$, implying greater sensitivity to tail probabilities in the unpacked condition and consistent with our earlier results.

6. CI Widths, Calibration, and Overall Forecasting Performance with Different Levels of Time Unpacking

In the previous sections, we show that time unpacking leads to wider confidence intervals and in general to assessments of greater uncertainty regarding a future uncertain quantity. This should lead to better calibration, as we know that assessors consistently tend to be overconfident. In this section, we discuss an experiment where it is possible to make observations of calibration for the assessed confidence intervals. We then investigate three interrelated issues.

One, we examine how different levels of time unpacking effect increases in CI widths. In the experiments presented in the previous sections, we have used two levels of unpacking. For example, in Experiment 1, participants in the unpacked condition made predictions for financial quantities for one month and two months ahead before finally making predictions for three months ahead. Here, we investigate the relationship between different levels of unpacking (i.e., one, two, three, and four levels) and the resulting increases in CI widths.

Two, we explore the relationship between increased CI widths from different levels of time unpacking and the corresponding changes in calibration. Do the increased CI widths, for example, lead to proportional improvements in calibration or perhaps to an overreaction that translates into *underconfidence*?

Three, we also consider another component of forecasting performance besides just calibration, namely, *sharpness*. The idea is that if we have two equally calibrated assessors, the better of the two then is the one who gives sharper assessments. For example, consider two weather forecasters assessing the probability of "rain tomorrow" in a region. One forecaster consistently gives the probability as 0.4, whereas the second forecaster gives a probability of 1 on 40% of the days and a probability of 0 on 60% of the days, and it rains only on the days when the second forecaster give a probability of 1. Both the weather forecasters in this case are perfectly calibrated, but the second one is *sharper* and more useful (see Murphy and Winkler 1977). Similarly, if two forecasters provide 90% confidence intervals and both are equally calibrated, the better of the two then is the one who tends to give tighter (and hence sharper) intervals. Now, consider another case where two forecasters are again providing 90% confidence intervals, one has a calibration of 60% and the other has 75%, but the first one consistently provides much narrower intervals than the second one. In this case, we need to make a trade-off between calibration and sharpness to judge the overall forecasting performance of the two assessors, and this is what we consider.

6.1. Experiment 5: Method

This experiment explores the time unpacking effect by using a simple random walk data-generating process. A total of 233 participants were invited to participate in this experiment. The problem given to the participants was a random walk starting at zero and with an equal chance of an increase or decrease of one unit every period. More formally, let X_t be the value of the process at time period t, with the starting value $X_0 = 0$. And, let $Z_t = X_t - X_{t-1}$, $t \geq 1$, be the increments in the process from time $t-1$ to t that can exclusively take the values -1 or $+1$. Additionally, it is assumed that the increments are independent and identically distributed Bernoulli trials with $P(Z_t = 1) = p = 0.5$ and $P(Z_t = -1) = 1 - p = 0.5$ for all t. Then, the random walk can be written as

$$X_t = X_0 + \sum_{k=1}^{t} Z_k = \sum_{k=1}^{t} Z_k, \quad t = 1, 2, \ldots . \tag{4}$$

If from time 0 to time t, the process increases by one unit r times, and thus decreases by one unit $t - r$ times, then

$$X_t = r - (t - r) = 2r - t, \quad 0 \leq r \leq t. \tag{5}$$

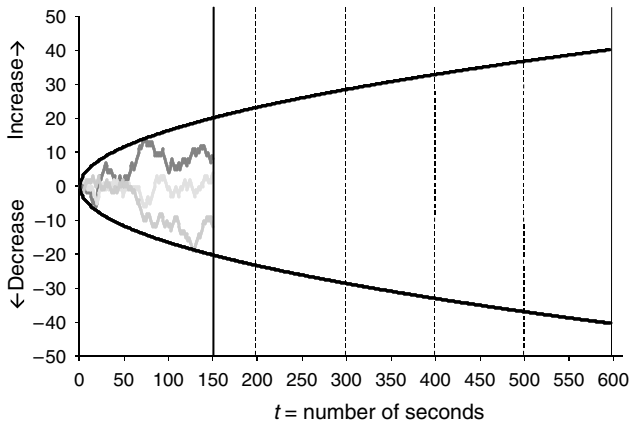
The random variable r is binomially distributed with t Bernoulli trials and $p = 0.5$. It follows that

$$\begin{aligned} E[X_t \mid t, p = 0.5] &= 2E[r \mid t, p = 0.5] - t \\ &= 2t(0.5) - t = 0, \quad \text{and} \end{aligned} \tag{6}$$

$$\begin{aligned} \text{Var}[X_t \mid t, p = 0.5] &= 2^2 \,\text{VAR}[r \mid t, p = 0.5] \\ &= 4t(0.5)^2 = t. \end{aligned} \tag{7}$$

The expected value of X_t is 0 at all t but the volatility of the process grows over time, with the variance increasing linearly in t. Further, for large t, the distribution of r can be approximated by a normal distribution with mean $0.5t$ and variance $0.25t$. Hence, for large t, the distribution of X_t can be approximated by a normal distribution with mean 0 and variance t. Figure 5 displays some sample random walks in t, along with the interval $0 \pm 1.645\sqrt{t}$. For large t, this interval should contain 90% of the realizations of the process. This is one of the simplest stochastic processes within the broad class of stochastic processes that we consider and describe in §2.

Figure 5 **Sample Paths and a 90% Confidence Interval for a Random Walk (Experiment 5)**



This data-generating process was described to the subjects (without the formal statistical representation) in terms of a repeated lottery game and 10 sample paths up to the 150th time period were uniquely generated for each participant. Time unpacking was manipulated between subjects with five conditions (*packed, unpacked one, unpacked two, unpacked three, and unpacked four*). Participants in the *packed* condition provided a 90% CI for the realization of the process at $t = 600$. Participants in the various unpacked conditions provided 90% CIs at the following time periods:

unpacked one: $t = 300$ and $t = 600$;

unpacked two: $t = 200$, $t = 400$, and $t = 600$;

unpacked three: $t = 300$, $t = 400$, $t = 500$, and $t = 600$;

unpacked four: $t = 200$, $t = 300$, $t = 400$, $t = 500$, and $t = 600$.

6.2. Experiment 5: Results

6.2.1. CI Widths. Table 5 shows the summary results for the 90% CIs for the 600th time period: means and standard errors of the means for lower bounds, upper bounds, and CI widths, along with the percentage increases in CI widths in each of the four unpacked conditions with respect to the packed condition and the marginal percentage increase in mean CI width from each additional level of unpacking. An analysis of variance (ANOVA) using CI width as a dependent measure and the five experimental conditions (packed, unpacked one, unpacked two, unpacked three, and unpacked four) as a between-subjects factor demonstrated a significant main effect of the conditions ($F(4, 228) = 4.71$, p(two-tailed) $= 0.001$). The mean CI widths in the unpacked one, unpacked two, unpacked three, and unpacked four conditions were greater than that in the packed condition by 30.83%, 42.33%, 55.03%, and 63.21%, respectively. Note that whereas the increase in mean CI width from just one level of unpacking was 30.83%, the marginal increase in mean CI width from additional levels of unpacking was 8.79% for unpacked two, 8.92% for unpacked three, and 5.27% for unpacked four. That is, the mean CI width seems to be increasing at a decreasing rate with the increase in unpacking levels, the most substantial and significant increase coming from just one level of unpacking with further marginal increases from additional levels of unpacking being statistically insignificant.

6.2.2. Implied Calibration. Because the data generating process is known in this case, we can obtain the *implied calibration* of the elicited CIs. If a stated 90% CI for time period t is (l, u) , $l \le u$, where l and u correspond to the ath and $a + b$th percentiles, respectively, in the distribution generated by the random walk at time t, then the implied calibration for the stated 90% CI is simply b%. For example, if the stated 90% CI for time period t corresponds to the 10th and the 90th of the distribution generated by the random walk at time t, then the implied calibration is $90\% - 10\% = 80\%$. In other words, a large number of trials in this case would lead to the assessed 90% CI capturing close to 80% of the observations. Perfect calibration for an assessed 90% CI would require an implied calibration of 90%.

Table 5 **90% CIs for $t = 600$ Days Ahead Under the Packed and Unpacked Conditions: Means and Standard Errors of the Means (SEM) for Lower Bounds (LB), Upper Bounds (UB), and CI Widths, Along with Percentage Increase in Mean CI Widths in the Unpacked Conditions vs. the Packed Condition and Marginal Percentage Increase in Mean CI Widths from Additional Level of Unpacking (Experiment 5)**

	Mean (SEM)			% increase in mean CI width: Unpacked conditions vs. packed condition	% marginal increase in mean CI width: Additional levels of unpacking
	LB	UB	CI width		
Packed ($N = 43$)	−21.63 (2.50)	20.74 (2.00)	42.37 (4.04)	—	—
Unpacked one ($N = 47$)	−27.00 (2.39)	28.44 (2.10)	55.44 (4.09)	30.83*	30.83*
Unpacked two ($N = 59$)	−30.50 (2.33)	29.81 (2.04)	60.31 (4.10)	42.33**	8.79
Unpacked three ($N = 43$)	−31.64 (2.86)	34.04 (2.69)	65.69 (5.18)	55.03**	8.92
Unpacked four ($N = 40$)	−32.72 (3.27)	36.44 (3.03)	69.15 (5.74)	63.21**	5.27

*$p < 0.05$; **$p < 0.01$ (two-tailed).

Table 6 Means and Standard Errors of the Means (SEM) for Implied Calibration (%) of the 90% CI at the 600th Time Period in a Random Walk, Along with Percentage Increase in Mean Calibration in the Unpacked Conditions vs. the Packed Condition and Percentage Increase in Mean Calibration from Additional Levels of Unpacking (Experiment 5)

	Mean (SEM) Calibration (%)	% increase in mean calibration: Unpacked conditions vs. packed condition	% marginal increase in mean calibration: Additional levels of unpacking
Packed ($N = 43$)	53.95 (3.55)	—	—
Unpacked one ($N = 47$)	66.27 (3. 71)	22.83*	22.83*
Unpacked two ($N = 59$)	69.93 (2.66)	29.61**	5.52
Unpacked three ($N = 43$)	71.62 (3.43)	32.75**	2.42
Unpacked four ($N = 40$)	73.92 (3.13)	37.01**	3.21

*$p < 0.05$; **$p < 0.01$ (two-tailed).

Table 6 shows the means and standard errors of the means for the implied calibration of the assessed 90% CIs at $t = 600$ in the random walk under the five experimental conditions, along with the percentage increase in the mean implied calibration in each of the unpacked conditions relative to the packed condition and the marginal percentage increase in calibration from each additional level of unpacking. An analysis of variance (ANOVA) using implied calibration as a dependent measure and the five experimental conditions as a between-subjects factor demonstrated a significant main effect of the conditions ($F(4, 228) = 5.69$, p(two-tailed) < 0.001). In all the conditions, the participants on average gave 90% CIs with calibration levels less than 90%, showing persistence of overconfidence. However, time unpacking reduced the degree of overconfidence. Consistent with the results of Table 5, calibration improved with each additional level of unpacking. However, the most substantial and only significant marginal improvement came from just one level of unpacking (22.83%), with further marginal improvements from additional levels of unpacking coming at a decreasing (and statistically insignificant) rate. Note, for example, from Tables 5 and 6, that just one level of unpacking leads to a 30.83% increase in the mean CI width, which then translates into a 22.83% improvement in calibration, whereas four levels of unpacking lead to a 63.21% increase in the mean CI width with a corresponding increase in calibration of only about 37%. This happens because to capture more of the tails of a distribution much larger increases in CI widths are needed.

6.2.3. Overall Forecasting Performance. Until now in this section, we have shown that the mean width of a CI increases with each additional level of time unpacking, but most of the increase comes with just one level of unpacking. These increases in mean CI widths translate into improved calibration, but the overconfidence does not disappear. Consistent with the results of mean CI widths, the improvement in calibration is most substantial and significant with just one level of unpacking and then flattens out rapidly with further additional levels of unpacking. But, what does this mean for some overall forecasting performance? In other words, do the benefits of time unpacking on calibration get overshadowed by worsening sharpness (i.e., increasing widths) of the assessed CIs? Scoring rules have been often proposed to provide ex ante incentives for careful and truthful assessments and ex post evaluation measures that provide a trade-off between calibration and sharpness of probabilistic judgments (see Winkler 1996 for a review). Although such scoring rules have been empirically studied for probabilistic judgments of binary events, they have not yet to the best of our knowledge been studied for estimation of fractiles or intervals.

Jose and Winkler (2009) recently proposed a framework for *strictly proper scoring rules* (i.e., scoring rules that provide ex ante incentives for truthful reporting) to evaluate quantile assessments and interval forecasts. Suppose that a stated $100(1-\gamma)$%CI for a random variable X_t, $t > 0$, is (l, u), $l \leq u$, where l and u correspond to $100\gamma/2$th and $100(1-\gamma/2)$th percentiles, respectively. If the realization of X_t is x, then the score is given by

$$S(\gamma, x) = 2 * g - \frac{(\gamma)(u-l)}{2} - (l-x)^+ - (x-u)^+, \quad (8)$$

where g is a scaling constant. In our analysis, without loss of generality, we set $g = 0$, and then the score is always negative. Because we are evaluating 90% CIs, $\gamma = 0.1$. Note that the scoring rule is piecewise linear, the penalty associated with the width of the interval is $(0.1)(u-l)/2$, and the penalty for the actual realization falling outside the interval is $l - x$ if $x < l$ (i.e., x is below the assessed lower bound) and is $x - u$ if $x > u$ (i.e., x is above the assessed upper bound). The overall score results from a trade-off between the penalty associated with the width of the interval and the penalty associated with x falling outside the interval. In the limiting case where an assessor provides a point forecast as the interval and that forecast equals the actual realization, the score is maximized at zero.

To compute scores for the assessed 90% CIs in our experiment, we first simulated 10,000 possible realizations from the underlying random walk distribution

Table 7 **Means and Standard Errors of the Means (SEM) for Average Score of the 90% CIs at the 600th Time Period in a Random Walk, Along with Percentage Increase in Mean Average Score in the Unpacked Conditions vs. the Packed Condition and Marginal Percentage Increase in Mean Average Score from Additional Levels of Unpacking (Experiment 5)**

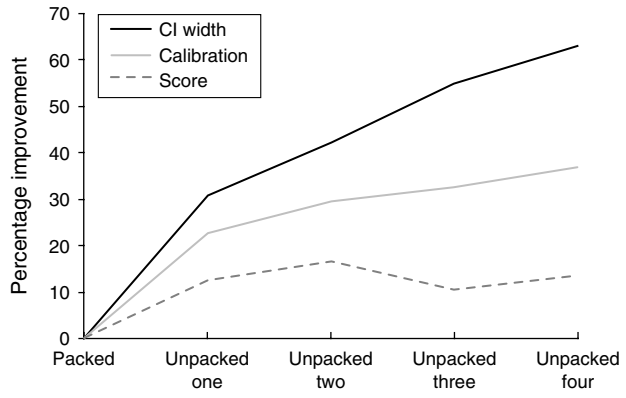
	Mean (SEM)	% increase in mean score (%)	
	Score	Unpacked conditions vs. packed condition	Additional level of unpacking
Packed ($N = 43$)	−5.76 (1.69)	—	—
Unpacked one ($N = 47$)	−5.04 (1.39)	12.49*	12.49*
Unpacked two ($N = 59$)	−4.81 (1.13)	16.54**	4.63
Unpacked three ($N = 43$)	−5.16 (1.16)	10.51	−7.23
Unpacked four ($N = 40$)	−4.99 (1.31)	13.37*	3.19

*$p < 0.05$; **$p < 0.01$ (two-tailed).

at $t = 600$. Based on these 10,000 realizations, we computed the *average score* for each participant. Table 7 shows the means and standard errors of the means for the average scores of the 90% CIs at $t = 600$ in the random walk under the five experimental conditions, along with the percentage increase in the mean average score in each of the unpacked conditions relative to the packed condition and the marginal percentage increase in mean average score from each additional level of unpacking. An analysis of variance (ANOVA) using this average score as a dependent measure and the five experimental conditions as a between-subjects factor demonstrated a significant main effect of the conditions ($F(4, 228) = 3.38$, p(two-tailed) $= 0.01$). We found that participants in all the four unpacked conditions had significantly higher scores than those in the packed condition. However, consistent with the results of CI widths and calibration in Tables 5 and 6, the most substantive (12.49%) and the only significant marginal improvement in the scores came from just one level of unpacking.

Figure 6 integrates the results of Tables 5–7, and shows the percentage increases in mean CI widths, mean calibration, and mean average score for overall forecasting performance for the different levels of time unpacking. There are significant benefits from just one level of unpacking. And, as the number of unpacking levels increase, the marginal benefits accrue at a decreasing rate.

Figure 6 **Percentage Improvement in Mean CI Width, Implied Calibration, and Average Score in the Unpacked One, Unpacked Two, Unpacked Three, and Unpacked Four Conditions vs. the Packed Condition (Experiment 5)**



To summarize, we show in this section that time unpacking leads to wider CIs (as in the experiments in previous sections), but most of this effect comes from just one level of time unpacking. The wider CIs translate into improved calibration. Overconfidence is not eliminated but is significantly reduced. Most of the improvement in calibration also comes from just one level of unpacking, with a decreasing marginal rate for further improvements from additional levels of unpacking. The improved calibration from wider CIs does not appear to be at the cost of overall forecasting performance that also brings into consideration the sharpness of assessments. In fact, a strictly proper scoring rule that provides a trade-off between calibration and sharpness suggests that overall forecasting performance also improves with time unpacking, with the most benefit coming from just one level of unpacking.

7. Discussion and Conclusion

Subjective probabilistic judgments are inevitable in prediction of uncertain future events in many real-life domains. For uncertain quantities that can take on a continuum of possible values, the most commonly used method for eliciting subjective probabilistic judgments is the *fractile method*, where either the fractiles or confidence intervals are assessed. It is well accepted in the extensive literature in this area that assessors consistently show overconfidence; that is, the subjective confidence intervals are much narrower than they should be. What this means is that when assessors are forecasting uncertain quantities into the future with, for example, 90% confidence intervals, their intervals are likely to capture the realized values much less than 90% of the time. In other words, the assessors are not well calibrated. Perfect calibration

by itself is not the only measure for the "goodness" of a probabilistic judgment (more on this below), but it still is very important. When the tail probability events have large consequences, for example, as in the most recent economic downturn and in many investment contexts, calibration is crucial. Also, in many risk management frameworks, where such assessments may be made at several levels of an analysis, overconfidence or overly narrow confidence intervals can quickly compound to a very serious underestimation of the overall uncertainty with possible enormous expected losses.

Attempts to reduce overconfidence in such cases have focused mainly on feedback and training, which have not yielded the promised results. One reason for this could be that the context in real life is continuously changing (such as the context for predicting oil prices or stock prices), and assessors, while being cognizant of the overconfidence bias, continue to fall prey to the thinking that "this time it is different."

In §2, we present a parsimonious model for overconfidence in subjective confidence intervals for a broad class of stochastic processes. If $\{X(t), 0 \leq t < \infty\}$ is a stochastic process, then we consider a class of stochastic processes in which the uncertainty about $X(t)$ is increasing in t. We specify this uncertainty by the width of the shortest prediction interval for $X(t)$, $t > 0$, which is increasing in t. This is a broad class of stochastic processes that includes, for example, processes widely applied to simulate prices or levels of financial assets. We then show that a compressed perception of time, where a time period t is perceived as $R(t)$, $R(t) < t$, leads to overconfident judgments or predictions of $X(t)$.

In §3, we describe a process, *time unpacking*, which consistently leads to wider subjective confidence intervals than is normally the case. Suppose we want to obtain an assessment in the form of a subjective confidence interval for an uncertain quantity, such as price of oil, three months ahead. Then, obtaining the confidence intervals for one month and two months ahead before doing so for three months ahead consistently leads to wider confidence intervals than those assessed directly for three months ahead. Of course, some assessors who are asked directly for three months ahead might be informally doing such time unpacking anyway in their construction of the three-month confidence intervals; making it explicit seems to be consistently effective. We show this not only with MBA students but also with professionals in the real-world forecasting in their domains of expertise. Herein lies one of the novelties (hopefully, among others) of this paper, we feel, where we do not focus just on experiments with students guessing answers to trivia general knowledge type of questions, as is the norm in many studies. Instead, we have looked at real-life analysts estimating uncertain quantities that they are used to and have to continuously track as part of their day to day business.

Why does this happen? In §4, we explore the psychological mechanisms of time unpacking that lead to assessments of greater uncertainty. As we show, in Experiment 3, this could be a manifestation of distortion of time perception based on the contrast effect. Three months when explicitly compared with one and two months might be perceived to be a longer time duration, and when compared explicitly with four and five months might appear to be a shorter time duration, in relation to the perceived time duration of three months without any explicit contrasts. A greater perceived time duration of a specific time period then appears to lead to consideration of greater possibilities for an uncertain quantity. Of course, it is entirely possible that some other psychological phenomenon might also be at play in the time unpacking format. Our focus and attention is primarily on the useful pragmatic manifestation of time unpacking, that it is a very simple process that consistently seems to yield wider subjective CIs that can reduce the degree of persistent overconfidence.

Furthermore, we show, in §5, that this process is robust to alternate elicitation methods. In Experiment 4a, instead to estimating confidence intervals respondents estimated the probability of an uncertain quantity being below or above a specific level at least once up to some point in the future (e.g., the probability of the Dow being below or above 15% its current level in the next three months). In Experiment 4b, respondents assessed the probability of an uncertain quantity being within a specified interval at some point in the future (e.g., the price of gold being within a defined interval three months ahead). In both cases, time unpacking leads to greater probabilities being assigned to the tails, which is consistent with wider confidence intervals. Again, we used not only students but also professionals in the financial industry forecasting uncertain quantities that are of continuous interest to them. This robustness further underlines the potential usefulness of time unpacking in terms of countering overconfidence in subjective probabilistic judgments.

How many levels of time unpacking should be done? Do the CI widths continue to increase at the same marginal rate with each additional level of time unpacking? Does this then lead to the other extreme, that is, *underconfidence* or assessments that are far too wide in terms of confidence intervals? How do any benefits of improved calibration from wider CIs play out in the overall forecasting performance that must also account for other aspects of assessments such as sharpness of forecasts? These are the types of questions we explore with Experiment 5 in §6.

We obtained assessments in the backdrop of a well-defined data generating process, which made it possible to compute the implied calibration of assessed confidence intervals. We used a random walk that was explicitly described to the subjects in terms of the data generating process, along with 10 random paths generated as an example uniquely for each subject. Subjects then assessed 90% confidence intervals for the realization of the stochastic process at a future time period under the conditions of packed time and of different levels of unpacked time. The mean CI width, when compared to packed time, increases with time unpacking with a decreasing marginal rate from each additional level of time unpacking. The most substantive and significant increase comes from just one level of time unpacking. The results are similar for calibration. Under all conditions, the 90% CIs shows calibration of less than 90%, reflecting persistence of overconfidence. However, calibration improves (i.e., overconfidence decreases) with time unpacking, the biggest improvement coming from just one level of unpacking with a decreasing rate of marginal improvement from additional levels of unpacking. In terms of overall forecasting performance, we consider a scoring rule for interval estimation that provides a trade-off between improved calibration and less sharpness. The improved calibration from time unpacking leads to better overall forecasting performance even when traded off against decreased sharpness. In other words, time unpacking leads to wider assessed intervals that are better calibrated, and the trade-off against the resulting reduced sharpness (increased width) of the intervals seems well worth as seen from the perspective of overall forecasting performance.

Admittedly, not all of our results show statistical significance. We have presented all our results as we observed, rather than only those that showed statistical significance. However, the empirical regularity of greater assessed uncertainty with unpacking across all the results gives us confidence that the unpacking heuristic can be a very useful practical tool not just to reduce the persistent problem of overconfidence but to also improve the overall forecasting performance. Furthermore, substantive benefits can be obtained from just one or two levels of time unpacking, thus not having to add substantially to the original task at hand. The unpacking should be done in a way that keeps the task "natural" within the context, or, in other words, the constituent time periods in the time unpacking should be constructed in a way that makes the cognitive task as convenient and natural as possible. For example, in some contexts, while looking one year ahead, it might be more natural to think in steps of quarters or six months than in some other unpacked way. We would be hesitant to claim that time unpacking is a "cure-all" for overconfidence, but feel that it can systematically reduce it. In that sense, it is a useful complement to any other efforts in reducing overconfidence.

Acknowledgments

The authors are extremely grateful to CLSA for sponsoring and enabling this project, especially to Chris Lobello who championed this project within CLSA and provided invaluable help in designing and running the experiments. The authors also thank Teck Ho (the department editor), Robin Hogarth, Jack Soll, Robert Winkler, the associate editor, and two anonymous referees for insightful comments at different stages of this paper. The Centre for Decision Making and Risk Analysis at INSEAD and the International Trading Institute at Singapore Management University provided financial support for this project.

References

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

Arkes HR, Christensen C, Lai C, Blumer C (1987) Two methods of reducing overconfidence. *Organ. Behav. Human Decision Processes* 39:133–144.

Ayton P (1997) How to be incoherent and seductive: Bookmakers' odds and support theory. *Organ. Behav. Human Decision Processes* 72:99–115.

Barberis N, Thaler R (2003) A survey of behavioral finance. Constantinides GM, Harris M, Stulz RM, eds. *Handbook of the Economics of Finance* (Elsevier, Amsterdam), 1052–1090.

Bearden JN, Wallsten TS, Fox CR (2007) Contrasting stochastic and support theory accounts of subadditivity. *J. Math. Psych.* 51:229–241.

Beebe-Center JG (1929) The law of affective equilibrium. *Amer. J. Psych.* 41:54–69.

Budescu DV, Du N (2007) The coherence and consistency of investors' probability judgments. *Management Sci.* 53:1731–1744.

Deaves R, Lüders E, Schröder M (2010) The dynamics of overconfidence: Evidence from stock market forecasters. *J. Econom. Behav. Organ.* 75:402–412.

Eisler H (1976) Experiments on subjective duration 1878–1975: A collection of power function exponents. *Psych. Bull.* 83:185–200.

Fox CR, Tversky A (1998) A belief-based account of decision under uncertainty. *Management Sci.* 44:879–895.

Griffin D, Brenner L (2004) Perspectives on probability judgment calibration. Koehler DJ, Harvey N, eds. *Blackwell Handbook of Judgment and Decision Making* (Blackwell, Malden, MA), 177–199.

Haran U, Moore DA, Morewedge CK (2010) A simple remedy for overprecision in judgment. *Judgment and Decision Making* 5:467–476.

Helson H (1964) Current trends and issues in adaptation-level theory. *Amer. Psychologist* 19:26–38.

Hogarth R (1975) Cognitive processes and the assessment of subjective probability distributions. *J. Amer. Statist. Assoc.* 70:271–289.

Hull JC (2008) *Options, Futures, and Other Derivatives* (Pearson, Upper Saddle River, NJ).

Johnson DDP, Fowler JH (2011) The evolution of overconfidence. *Nature* 477:317–320.

Jose VRR, Winkler RL (2009) Evaluating quantile assessments. *Oper. Res.* 57:1287–1297.

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

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

Lichtenstein S, Fischhoff B, Phillips LD (1982) Calibration of probabilities: The state of the art to 1980. Kahneman D, Slovic P, Tversky A, eds. *Judgment Under Uncertainty: Heuristics and Biases* (Cambridge University Press, Cambridge, UK), 305–334.

Manstead ASR, Wagner HL, MacDonald CJ (1983) A contrast effect in judgments of own emotional state. *Motivation and Emotion* 7:279–290.

McDonald RL (2005) *Derivatives Markets* (Pearson/Addison-Wesley, Boston).

Morgan MG, Henrion M, Small M (1990) *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis* (Cambridge University Press, Cambridge, UK).

Murphy AH, Winkler RL (1977) Reliability of subjective probability forecasts of precipitation and temperature. *Applied Statist.* 26:41–47.

Ross SM (1983) *Stochastic Processes* (John Wiley & Sons, New York).

Russo JE, Schoemaker PJH (1992) Managing overconfidence. *Sloan Management Rev.* 33(2):7–17.

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

Stevens SS (1975) *Psychophysics: Introduction to Its Perceptual, Neural and Social Prospects* (John Wiley & Sons, New York).

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

Tversky A, Koehler DJ (1994) Support theory: A nonextensional representation of subjective probability. *Psych. Rev.* 101:547–567.

Winkler RL (1996) Scoring rules and the evaluation of probabilities. *Test* 5:1–60.

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