

Overconfidence in interval estimates: What does expertise buy you? ☆

Craig R.M. McKenzie^{a,*}, Michael J. Liersch^b, Ilan Yaniv^c

^a *Rady School of Management and Department of Psychology, University of California, San Diego, 9500 Gilman Drive, MC 0553, La Jolla, CA 92093-0553, USA*

^b *Department of Psychology, University of California, San Diego, 9500 Gilman Drive, MC 0553, La Jolla, CA 92093-0553, USA*

^c *Hebrew University, Jerusalem, Israel*

Received 14 July 2005

Available online 28 March 2008

Accepted by Robyn Dawes

Abstract

People's 90% subjective confidence intervals typically contain the true value about 50% of the time, indicating extreme overconfidence. Previous results have been mixed regarding whether experts are as overconfident as novices. Experiment 1 examined interval estimates from information technology (IT) professionals and UC San Diego (UCSD) students about both the IT industry and UCSD. This within-subjects experiment showed that experts and novices were about equally overconfident. Experts reported intervals that had midpoints closer to the true value—which increased hit rate—and that were narrower (i.e., more informative)—which decreased hit rate. The net effect was no change in hit rate and overconfidence. Experiment 2 showed that both experts and novices mistakenly expected experts to be much less overconfident than novices, but they correctly predicted that experts would provide narrower intervals with midpoints closer to the truth. Decisions about whether to consult experts should be based on which aspects of performance are desired.

© 2008 Elsevier Inc. All rights reserved.

Keywords: Overconfidence; Expertise; Interval estimates

People often express uncertain values in terms of an interval, such as when they estimate their arrival time (“Between 5:00 and 5:30”), another person's age (“35 to 40”), or next year's inflation rate (“3% to 5%”). The accuracy of such estimates is usually measured in terms of hit rate: How often do the intervals contain the true value? Hit rates are often compared to the degree of confidence reported in the intervals. For example, participants might be asked to report low and high values for the populations of various cities such that they are 90% confident that each

resulting interval contains the city's true population. If people were well calibrated, 90% of their 90% confidence intervals would contain the true value. However, true values typically fall within such intervals between 30% and 60% of the time, indicating extreme overconfidence (e.g., Alpert & Raiffa, 1982; Juslin, Wennerholm, & Olsson, 1999; Klayman, Soll, González-Vallejo, & Barlas, 1999; Lichtenstein, Fischhoff, & Phillips, 1982; Soll & Klayman, 2004; Teigen & Jørgensen, 2005; Yaniv & Foster, 1997).¹

☆ This research was supported by National Science Foundation Grant SES-0551225 and by Israel Science Foundation Grant 344/05. Some of the results were presented at the 2004 Annual Meeting of the Society for Judgment and Decision Making in Minneapolis, MN.

* Corresponding author. Fax: +1 858 534 7190.

E-mail address: cmckenzie@ucsd.edu (C.R.M. McKenzie).

URL: <http://psy.ucsd.edu/~mckenzie/> (C.R.M. McKenzie).

¹ Although we will usually refer to intervals associated with a particular level of confidence as “subjective confidence intervals” (or “confidence intervals” for short), readers should be aware that these intervals are sometimes referred to in the literature as “credible intervals”, “uncertainty intervals”, “probabilistic prediction intervals”, and “fractile assessments” (Teigen & Jørgensen, 2005).

Attempts to overcome overconfidence in interval estimates have been only modestly successful, indicating that the degree of overconfidence is not only large, but also robust (Alpert & Raiffa, 1982; Lichtenstein et al., 1982). Such overconfidence has practical, as well as theoretical, significance. Russo and Schoemaker (1992) described a leading US manufacturer that elicited a projected range of sales from its marketing staff in order to plan the production capacity of a new factory. The range turned out to be too narrow, and the new factory was incapable of meeting the unexpected demand.

Most confidence interval studies ask undergraduate students about values they are unlikely to know much about (e.g., “What is the gestation period of an Asian elephant?”), and one might wonder whether being more knowledgeable reduces overconfidence. Several studies have examined how well experts assign probabilities to events, and the results have been mixed (e.g., Christensen-Szalanski & Bushyhead, 1981; Keren, 1987; Lichtenstein & Fischhoff, 1977; Murphy & Winkler, 1977; Oskamp, 1965; for reviews, see Camerer & Johnson, 1997; Koehler, Brenner, & Griffin, 2002). Only a few studies have examined experts’ interval estimates, and those results have been mixed as well. Russo and Schoemaker (1992) asked advertising, petroleum, and money management professionals (among others) for interval estimates in their domain of expertise, and these experts’ 90% and 95% confidence intervals were typically accompanied by hit rates of 40–60%. Although these results seemed to suggest that experts are just as overconfident as novices, no direct comparison was made between them. Önkal, Yates, Simga-Mugan, and Öztin (2003) did make direct comparisons by studying experts’ and (sophisticated) novices’ ability to predict foreign exchange rates. The experts tended to outperform the novices in terms of point predictions and predicting direction of change, but there were no differences in hit rates for 90% confidence intervals. When predicting exchange rates 1 day and 1 week ahead, hit rates for the two groups ranged between 40% and 56%.

Yates, McDaniel, and Brown (1991) found that graduate students in finance classes (“semi-experts”) were *more* overconfident than undergraduate finance students when predicting changes in stock prices, suggesting that expertise can even exacerbate overconfidence (see also Staël von Holstein, 1972). However, these participants assigned probabilities to six fixed, nonoverlapping intervals for each stock (“increase in price greater than 10%”, “increase in price between 5% and 10%”, and so on), which is different from generating a single (high) confidence interval. Furthermore, although there was a reliable difference in accuracy, the graduate and undergraduate students nonetheless provided very similar predictions.

Finally, Tomassini, Solomon, Romney, and Kroghstad (1982) studied professional auditors’ subjective probability distributions for financial statement account balances. They concluded that these experts tended to be

underconfident, but only for relatively low-confidence intervals (50% and 80% intervals). For high-confidence intervals (98%), the auditors were overconfident, but much less so than the typical findings in the literature (they had hit rates of 93% for their 98% confidence intervals). However, there was no control (i.e., novice) condition, the participants reported multiple fractiles (rather than a single high-confidence interval), and they were given “detailed training” by the experimenters about subjective probability distributions (and given opportunities to change their responses after further review and consultation with the experimenters), making it difficult to know if their auditing expertise was responsible for the diminished overconfidence.

In short, although it is clear that overconfidence in (high) confidence interval estimates is large, robust, and can have important consequences, little is known about differences between experts and novices. It seems safe to say that experts are overconfident, but it is unclear how they compare to novices, and knowing how they compare is important. If experts are just as overconfident—or more overconfident—than novices, why consult them at all?

In this article, we investigate three related topics. First, we examine hit rates for high-confidence intervals provided by information technology (IT) professionals and University of California, San Diego (UCSD) undergraduate students to questions about both the IT industry and UCSD. Thus, for questions about UCSD, UCSD students are experts and IT professionals are novices, while for questions about the IT industry, expertise is reversed. This within-subjects examination of expertise and interval estimates is the first of its kind, as far as we know, and allows us to make direct comparisons between expert and novice hit rates while controlling for differences between the groups of participants and the sets of questions. The “naïve” prediction is that experts will be less overconfident simply because they know more, but, as mentioned, there is little evidence to support this prediction and, depending on how relevant the Yates et al. (1991) study is perceived to be, there is even evidence supporting the opposite prediction.

Second, we examine more than just hit rates by evaluating both interval width and error, where error is defined as the absolute distance between the midpoint of an interval and the true value (Yaniv & Foster, 1995, 1997). Such measures can provide insight into why hit rates are relatively high or low. For example, wider intervals will generally increase hit rate, all else equal. If experts have higher hit rates than novices, it may be because they know more about the limits of their knowledge (i.e., have more metaknowledge) and therefore provide wider intervals. But intervals better centered on the truth will also lead to higher hit rates, holding interval width constant. Instead of reporting wider intervals, experts may make more accurate point predictions (e.g., Önkal et al., 2003), or best guesses of the true value. This could lead

their interval midpoints to be closer to the truth and increase hit rate. Of course, experts might report intervals that are both wider *and* better centered on the truth, which would increase hit rate even more.

But what if experts and novices tend to have the same hit rates? This would occur, of course, if experts and novices report intervals with similar widths and errors. Another possibility, though, is that, relative to novices, experts report intervals that are narrower (lowering hit rate), but have midpoints closer to the truth (increasing hit rate). The net effect could be that hit rate, and hence overconfidence, remains unchanged.

Because the emphasis in the confidence interval literature is on increasing hit rates, the discussion tends to focus on how to get people to widen their intervals. That is, wider intervals are usually considered better intervals. However, experts might provide narrower intervals because they are more informative. Consider two estimates of the selling price of a house. One estimate is \$480,000–\$520,000, and the other is \$200,000–\$800,000. The correct value turns out to be \$500,000. Although both intervals contain the true value and have the same midpoint, the narrower interval would be much more informative (e.g., if one were trying to decide on a bid or asking price). Yaniv and Foster (1995, 1997) have shown that, when both generating and evaluating intervals, people are concerned with accuracy *and* informativeness. Indeed, if the actual selling price had turned out to be \$525,000—outside the narrow interval, but inside the wide one—many people would probably still consider the narrow interval to be superior because it is so much more informative. (For a Bayesian interpretation of results like these, see McKenzie & Amin, 2002.) In short, hit rate is not all that matters when people evaluate intervals. Thus, even if experts do not have higher hit rates than novices, they might nonetheless provide narrower, and hence more desirable, intervals, as well as intervals with less error. Indeed, it would be noteworthy if experts were able to provide relatively narrow intervals without sacrificing hit rate.

Evidence suggesting that experts might provide narrower intervals that are better centered on the truth—but with no improvement in hit rate—was provided by Yaniv and Foster (1997), who found that as the midpoints of (novice) judges' intervals were closer to the truth, their intervals got narrower. As a result, hit rate remained largely constant. To the extent that midpoints closer to the truth are indicative of greater knowledge, the result suggests that expertise might manifest itself in the manner described above. Importantly, though, Yaniv and Foster studied only novices. They showed only a correlation between interval width and midpoint accuracy, and they assumed that greater midpoint accuracy indicated greater knowledge. Because we manipulate levels of knowledge (expertise) in Experiment 1, we can examine the causal role of domain knowledge in determining interval width

and midpoint accuracy and the resulting net effect on hit rate. We hasten to add, however, that Önkál et al. (2003) not only found no difference between hit rates for experts and novices predicting foreign exchange rates, they also found no differences between interval widths. Thus, in contrast to Yaniv and Foster's (1997) suggestive results using only novices, Önkál et al. (2003) found no evidence of a relationship between knowledge and interval size comparing experts and novices. In short, it is difficult to predict how expertise will manifest itself in interval estimates, but Experiment 1's within-subjects design should reveal differences, if they exist.

In Experiment 2, we examine our third and final topic: People's expectations of experts' confidence intervals. For example, do people expect higher hit rates from experts? If experts are just as overconfident as novices, consulting experts for better calibrated intervals would be pointless. However, perhaps experts report tighter intervals better centered on the truth—and this is what people want. In this case, consulting experts would be worthwhile. Interval estimates can be evaluated on multiple dimensions—hit rate, informativeness, and error—and we are interested in examining which dimensions experts excel at (if any) and comparing their actual performance to what others expect of them.

An issue in studies such as this one is how to operationalize expertise. In the current case, we simply sampled from two different populations—IT professionals and UCSD undergraduate students—and asked them to provide interval estimates of quantities regarding both the IT industry and UCSD. It suffices for our purposes if IT professionals are more knowledgeable than UCSD students about the IT industry, and UCSD students are more knowledgeable than IT professionals about UCSD. The task itself provides a manipulation check. Because Önkál et al. (2003) found that experts reported intervals with midpoints closer to the truth, we should find this as well. This would indicate that our task is tapping into domain differences of the kind we are interested in.

Although the experts in our studies have greater domain knowledge relative to the novices, we do not ask the experts to perform tasks they typically perform (e.g., Önkál et al., 2003; Tomassini et al., 1982). Our studies ask experts about topics they are relatively knowledgeable about but do not necessarily make judgments about on a regular basis (if at all; e.g., Russo & Schoemaker, 1992). For example, for the question, “In 2001, what percentage of UCSD students were from San Diego?”, a UCSD student would be more knowledgeable about the topic than the typical IT professional, although this would not be the sort of judgment a UCSD student regularly makes. Thus, our studies examine the role of domain knowledge, not process knowledge, in determining interval estimates.

Another issue when studying expert–novice differences is how to control for differences between the two

populations (e.g., age, education). Our within-subjects design provides the control. Because we ask the IT professionals and the UCSD students about both the IT industry and UCSD, effects of expertise are revealed by interactions, not main effects, and therefore differences between the populations will matter little.

In sum, Experiment 1 examines whether and why experts and novices differ in terms of hit rates. Expertise could manifest itself in different ways when reporting interval estimates, and it may or may not lead to less overconfidence. Experiment 2 examines whether people's expectations about expert performance are accurate.

Experiment 1

Method

Participants

Participants were 92 UCSD students (24% males and mean age of 20.5 yr) and 43 IT professionals (65% males and mean age of 37.4 yr). UCSD students participated in exchange for partial credit in psychology courses whereas IT professionals volunteered their time. One of the authors had access to IT professionals through prior work experience and mailed surveys to 50 of them, 43 of whom completed and returned the survey. The IT professionals were spread throughout the USA (one lived in Canada), but most (34) lived in California. They were (or recently had been) employed by various leading (e.g., IBM, Hewlett-Packard, Deloitte Consulting) and start-up (e.g., RightWorks/i2, BroadBand Office) companies. The modal IT professional was employed by Hewlett-Packard. All participants completed the survey during the fall of 2002.

Procedure

UCSD students filled out the survey in a laboratory setting. They read that they would be presented with 40 questions (see Appendix A) asking about values that they would probably be uncertain about. They were to write down a low and a high value such they were 90% confident that the true value fell within the resulting interval. A completed example was provided that asked about Bill Gates' net worth. To minimize errors, participants were instructed to write the words "million," "billion," or "trillion" for answers of that magnitude, but to write the actual number for answers less than a million. They were further told that they should expect about 10% misses, or four out of 40 questions, for 90% confidence intervals, and that none of the questions was meant to be "tricky;" they should interpret any question they felt was ambiguous in the most natural way.

Because IT professionals did not complete the survey in the laboratory, there were some additional instructions for them. In addition to being provided with the

necessary contact information, they were instructed not to look up answers to the questions, not to talk to others about the questions while completing the survey, and to complete the survey in one sitting.

After reporting 90% confidence intervals for the 40 questions, all participants answered some demographic questions. We also asked the UCSD students if they had ever worked in the IT industry (none had) and asked the IT professionals if they had ever attended UCSD (one had).

Design

In addition to there being two groups of participants (UCSD students and IT professionals), there were two question domains. Twenty of the questions concerned UCSD and 20 concerned the IT industry. Thus, UCSD students were considered experts for the UCSD questions and novices for the IT questions, whereas the opposite was true for the IT professionals. Furthermore, half of the questions for each domain were "bounded," asking for a percentage (e.g., "In 2001, what percent of UCSD students were graduate students?"), and half were "unbounded," asking for a quantity ("How many undergraduates were enrolled at UCSD in 2001?"). Finally, half of the participants were presented with the questions in a predetermined random order, and half were presented with the reverse order.

Dependent measures and predictions

Hit rate was determined for each participant for a set of questions by dividing the number of times the participant's interval contained the true value by the number of questions. If the true value equaled an interval boundary, it was counted as a hit. This is the typical measure used to evaluate interval estimates (e.g., Klayman et al., 1999; Lichtenstein et al., 1982; Russo & Schoemaker, 1992; Soll & Klayman, 2004). As mentioned, though, it is unclear whether expert and novice hit rates will differ.

In an attempt to understand why the hit rates of experts and novices do or do not differ, we used two additional measures (inspired by Yaniv & Foster, 1995, 1997).

Error equaled $|t - m|$, where t is the true value and m is the midpoint of a given interval, or $(\text{high value} + \text{low value})/2$.² Holding interval size constant, reducing error will generally increase hit rate. We expected experts to

² Using the midpoints of intervals to assess error is reasonable because doing so leads to similar results when compared to using participants' median estimates (50th fractiles; Soll & Klayman, 2004) and when compared to using participants' "best guesses" (Yaniv & Foster (1997)). Thus, midpoints are good proxies for participants' median or best estimates. This might not always be the case, though, such as when estimates are close to a natural boundary (e.g., zero) and the intervals are asymmetric. However, we suspect that such cases are rare and have limited or no bearing on our conclusions, as the earlier results suggest.

have lower error (Önkal et al., 2003; see also Yaniv & Foster, 1997). Assuming experts have lower error, interval size will determine whether experts and novices differ in terms of hit rate.

Interval size, g , equaled high value – low value, and is a measure of informativeness (narrower intervals are more informative). Holding error constant, widening intervals will generally increase hit rate. Thus, experts could have higher hit rates than novices due to reduced error and/or increased interval size. However, because Yaniv and Foster (1997) found that participants with lower error (who are presumably more knowledgeable) tended to report narrower intervals, experts might report narrower intervals as well (but see Önkal et al., 2003). If experts have both lower error and report narrower intervals, this would mean that their intervals are both better centered on the truth and more informative. However, whether their hit rates are better or worse than novices' will depend on how much experts narrow their intervals relative to how much they reduce their error. That is, it is the relationship between error and interval size that is key. Conceptually speaking, if experts narrow their intervals more (less) than they reduce their error, they will have lower (higher) hit rates than novices. And if they narrow their intervals to the

same extent that they reduce error, expert and novice hit rates will be the same.

Results

We present the results separately for the bounded (percentage) and unbounded (quantity) questions because (a) the dependent measures (except for hit rate) are on different scales and have different distributions, and (b) as will be seen, the pattern of results was somewhat different for the two question types. In addition, eliminating the one IT professional who had previously attended UCSD had virtually no effect on the results and was included in the analyses. Finally, of the 5400 requested intervals (135 participants \times 40 questions), 240 (4%) were unusable. Of these 240 intervals, 215 (90%) were unusable because participants either reported a quantity when asked for a percentage or reported a percentage when asked for a quantity.

Because our samples of participants and questions differ on a number of dimensions (e.g., age and education in the case of participants), main effects are of little interest. Interactions, on the other hand, reveal effects of expertise and are not easily explained in terms of differences between groups or questions.

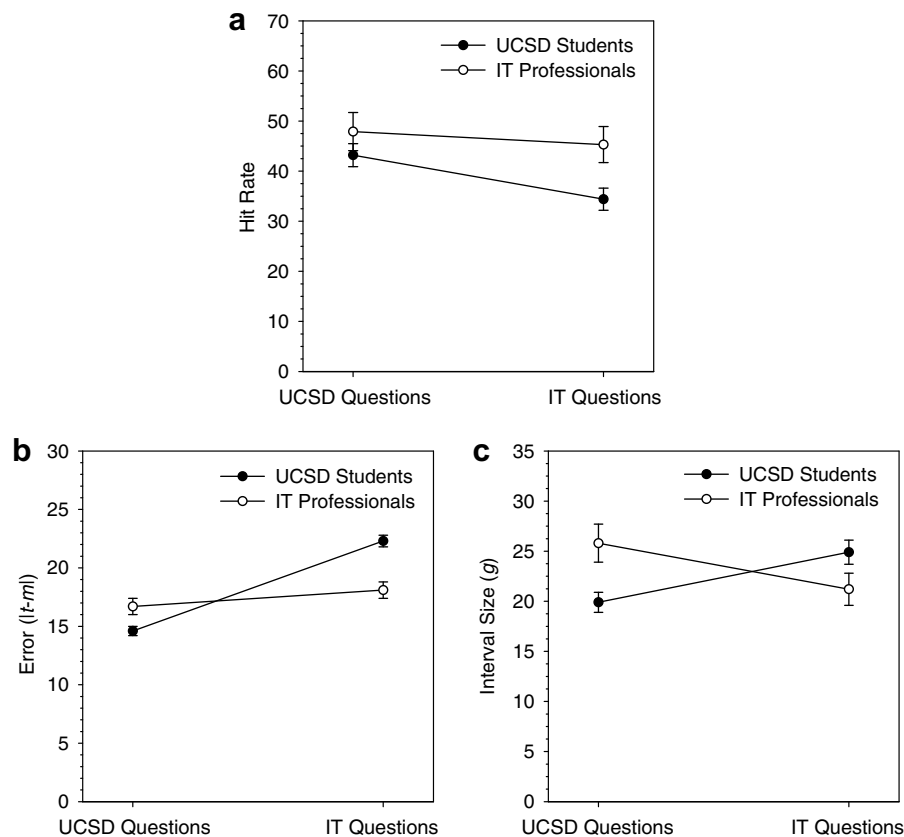


Fig. 1. Experiment 1 results for the bounded (percentage) questions. Standard error bars are shown.

Bounded (percentage) questions

The mean results for the bounded questions are shown in Fig. 1. Fig. 1a shows hit rate for UCSD students and IT professionals in both domains (UCSD and IT). The overall hit rate was 41% for these 90% confidence intervals, replicating previous findings of extreme overconfidence in interval estimates. IT professionals had a higher hit rate than UCSD students in both domains, and both groups had a higher hit rate for UCSD questions than for IT questions. A 2 (Group: UCSD students, IT professionals) \times 2 (Domain: UCSD, IT) mixed-model ANOVA on hit rate revealed main effects of Group ($F(1,133) = 4.4$, $p = .038$) and Domain ($F(1,133) = 9.52$, $p = .003$). Importantly, the interaction was not significant ($p = .094$), showing that participants did not have reliably higher hit rates in their respective areas of expertise. Hit rates for experts and novices were 44% and 41%, respectively.

The results for error, $|t - m|$, are shown in Fig. 1b. There was no effect of Group ($p = .14$), but there was an effect of Domain ($F = 69.9$, $p < .001$), with UCSD questions resulting in smaller error. (All analyses were 2×2 mixed-model ANOVAs as above.) Most important is that the interaction was significant ($F = 33.0$, $p < .001$): experts had lower error than novices. UCSD

students had less error than IT professionals for UCSD questions, but more error than IT professionals for IT questions.

Fig. 1c shows the results for interval size (g) and reveals why experts had lower error than novices but nonetheless similar hit rates: experts' intervals were narrower. UCSD students provided smaller intervals than IT professionals for UCSD questions, but they provided wider intervals than IT professionals for IT questions. Only the interaction was significant ($F = 70.6$, $p < .001$).

Unbounded (quantity) questions

The mean results for the unbounded questions are shown in Fig. 2. Fig. 2a shows hit rates. The overall hit rate was 39%, again replicating previous findings of extreme overconfidence. There was no effect of Group ($p = .16$), but there was an effect of Domain ($F(1,133) = 81.2$, $p < .001$), with higher hit rates for UCSD questions. There was also an interaction ($F(1,133) = 14.8$, $p < .001$): UCSD students had a slightly higher hit rate than IT professionals for UCSD questions, but they had a considerably lower hit rate than IT professionals for IT questions. Thus, for these questions, experts had higher hit rates (43% vs. 36%), which is consistent with the “naïve” prediction that experts would be less overconfident.

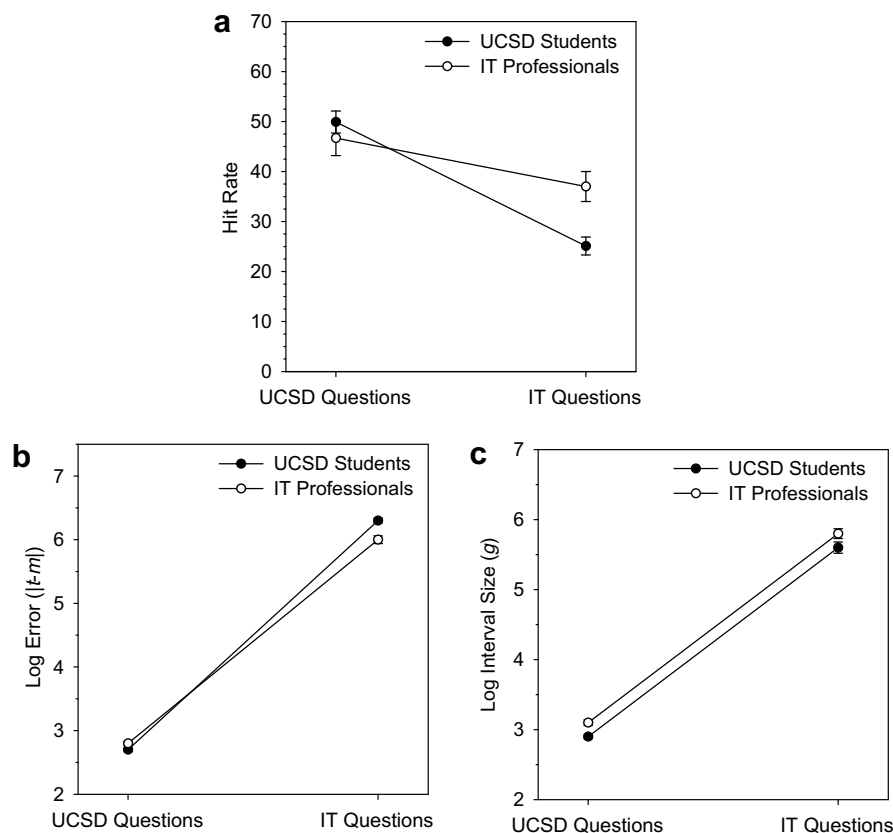


Fig. 2. Experiment 1 results for the unbounded (quantity) questions. Standard error bars are shown.

Log (base 10) error ($|t - m|$) is shown in Fig. 2b. (Logs were taken of error and interval size because of the skewed distributions for the unbounded questions.) There was no effect of Group ($p = .2$), but there was a large effect of Domain ($F = 8630.8$, $p < .001$), with IT questions resulting in greater error. As with the bounded questions, the interaction was also significant ($F = 37.6$, $p < .001$): UCSD students had lower error than IT professionals for UCSD questions, but they had greater error than IT professionals for IT questions. Expertise again resulted in interval midpoints closer to the truth.

Fig. 2c shows log interval size (g). There was an effect of Group, with UCSD students providing smaller intervals than IT professionals ($F = 6.3$, $p = .013$). There was also an effect of Domain, with smaller intervals reported for UCSD questions than for IT questions ($F = 2360.8$, $p < .001$). Most important, and unlike the bounded questions, the interaction for these unbounded questions was not significant ($F < 1$); that is, experts did not report smaller intervals. Although UCSD students provided smaller intervals than IT professionals for UCSD questions, they also provided smaller intervals for the IT questions. This, in turn, accounts for the effect of expertise on hit rate for the unbounded questions: UCSD students had greater error on IT questions and provided *smaller* intervals for these questions.

Discussion

At issue was whether and why experts and novices differ in terms of hit rates when providing interval estimates. For the bounded (i.e., percentage) questions, experts and novices had similar hit rates (44% vs. 41%, Fig. 1a) and thus a similar (large) degree of overconfidence. Importantly, however, it was also revealed that, despite the similar hit rates, experts had both lower error ($|t - m|$; Fig. 1b) and provided narrower (g ; Fig. 1c), or more informative, intervals. Reducing error increases hit rate, while narrowing intervals decreases it. The net effect was that the experts and novices had similar hit rates. It is of considerable interest that people in general and experts in particular appear able, at least under these conditions, to reduce both error and interval width so as to maintain a somewhat constant hit rate.

The results for the unbounded questions showed a slightly different pattern. In fact, the results regarding hit rate (Fig. 2a) were consistent with the “naïve” prediction: experts had higher hit rates (albeit modestly so; 43% vs. 36%). Measures of error and interval size help explain why. Relative to the bounded questions, the anomalous result appears to be that UCSD students had an especially low hit rate for IT questions. The pattern of results for error (Fig. 2b) is similar to that for the bounded questions (i.e., experts had lower error for both types of question), so differences in error do not account for the different hit rate result. Instead, the explanation

appears to lie in the results for interval size: there was an effect of expertise on interval size for bounded, but not unbounded, questions. For unbounded questions, UCSD students provided smaller intervals than IT professionals for UCSD questions—as they did for the bounded questions—but they also provided smaller intervals for IT questions (Fig. 2c), which is different from the results for the bounded questions. Thus, for the IT questions, UCSD students, relative to IT professionals, had more error and provided *smaller* intervals. Though the difference in interval size is not large (but keep in mind that $\log g$ is shown in Fig. 2c), it is the *combination* of larger error and smaller intervals that accounts for the decreased hit rate for UCSD students answering IT questions.

Although the difference between expert and novice overconfidence differed only modestly between the two question types, it is nonetheless of interest to speculate why. We suspect that it is because the IT questions were of such a large magnitude that it was difficult for UCSD students to provide sufficiently wide intervals. The mean true value for the unbounded IT questions was over 27 billion, whereas the mean true value for the unbounded UCSD questions was only about 12,000. When dealing with such large values, novices might have difficulty balancing interval width and error, which people seem to do so well otherwise. In this case, the result was a decreased hit rate. This account predicts that the differences will disappear when mean true value is controlled for. Indeed, the bounded questions essentially controlled for this factor (the mean true values for the bounded UCSD and IT questions were 36 and 38, respectively), and expertise had no effect on hit rate for these questions. We suspect, then, that the different hit rates for experts and novices for the unbounded questions is the exception, not the rule, and that the typical finding will be similar hit rates for experts and novices. We will not pursue this further because the difference between the hit rate results for the bounded and unbounded questions was not large and we do not think it is theoretically important. Instead, our two subsets of questions fortuitously revealed different ways that expertise can benefit interval estimates.

Our two samples of participants, IT professionals and UCSD students, differed on a number of dimensions, including age, income, education, and gender. Klayman et al. (1999) found that males were more overconfident than females when reporting interval estimates. Because our IT professionals had a much higher proportion of males, this is a potentially important confound in our study. However, we found that the IT professionals were less overconfident overall. More important is that our experimental design mitigated such confounds. Although main effects of Group (and Domain for that matter) must be interpreted with care, our focus was on expertise, any effects of which were revealed by inter-

actions, not main effects. Therefore, effects of expertise are not easily explained in terms of differences between the two samples (or question domains).

Experiment 2

What do people expect from experts who provide interval estimates? If they expect intervals with a high probability of containing the true value, or even much better calibration than novices, they are likely to be dissatisfied. But if people look to experts for intervals that are relatively informative or centered on true values, then consulting experts might be worthwhile. Thus, whether people make good decisions about consulting experts for interval estimates depends on whether they know what experts are and are not good at.

In Experiment 2, we asked UCSD students to predict the results of Experiment 1 for the bounded (percentage) questions. Thus, the participants in Experiment 2 made predictions about both experts and novices who were both similar to and different from themselves, depending on the target group (UCSD students or IT professionals) and the domain (UCSD or the IT industry) asked about.

Other studies have asked participants to estimate how many of their interval estimates contained the true value, and a consistent finding has been that, for $X\%$ confidence intervals, people's estimates fall somewhere between $X\%$ and their actual hit rate (Soll & Klayman, 2004; Teigen & Jørgensen, 2005; see also Snizek & Buckley, 1991). Indeed, Önköl et al. (2003) found this result for both experts and novices. However, our Experiment 2 differs in important ways. We asked participants to predict how *others* will perform and, in particular, how they expect experts to perform compared to novices. Moreover, participants reported their expectations about expert–novice differences with respect to interval width and interval error in addition to hit rate.

Method

Participants were 136 UCSD students (32% males and a mean age of 20.2 yr) who received partial course credit. During the spring of 2006, they filled out a questionnaire asking them to predict some of the results of Experiment 1. After reading the instructions (Appendix B), they were presented with the 10 bounded (percentage) UCSD questions and the 10 bounded IT questions used in Experiment 1. It was made clear that they were not to report confidence intervals; they were simply to review the task performed in Experiment 1. We used only the bounded questions because they were all on the same scale, which simplified our asking questions about measures such as interval width and error, as we describe below.

After looking over the 20 bounded questions from Experiment 1, the current participants were then asked several questions. For expected hit rate, participants were asked how many of the UCSD students' 90% confidence intervals, on average, contained the true value for the 10 UCSD questions and, separately, for the 10 IT questions. They then repeated the task, but estimated how many of the IT professionals' intervals contained the true value for both sets of questions.

They were also asked who they thought reported wider intervals (where width was defined as the high value minus the low value) for the UCSD questions, on a scale of -3 (UCSD students reported much wider intervals) to 3 (IT professionals reported much wider intervals). The task was repeated for the IT questions.

Finally, the participants considered interval error, defined as the absolute difference between the middle of an interval and the true value. They reported who they thought, on average, reported intervals with greater error for UCSD questions and, separately, for IT questions, both using scales ranging from -3 (UCSD students reported intervals with much greater error) to 3 (IT professionals reported intervals with much greater error).

Note that participants provided four responses to the hit rate questions, but only two responses to the width and error questions. This was because the width and error questions asked only for a comparison between experts and novices rather than separate estimates for each. We thought it best to ask for judgments of relative, rather than absolute, width and error because that would be easier for participants and lead to more interpretable responses.

Half of the participants answered the questions in the order above (hit rate, interval width, interval error), and half answered them in the reverse order.

Results

Expected hit rates

We conducted a 2 (Target Group: UCSD students, IT professionals) $\times 2$ (Question Domain: UCSD questions, IT questions) $\times 2$ (Question Order) mixed-model ANOVA on expected hit rates, using the first two variables as within-subjects variables. There were two main effects, one of which was Target Group, $F(1, 134) = 11.8$, $p < .001$. The UCSD participants expected IT professionals to have higher hit rates than the UCSD students (6.3 vs. 6.0 out of 10). (Interestingly, IT professionals did have overall higher hit rates in Experiment 1.) The other main effect occurred for Question Order, $F(1, 134) = 3.9$, $p = .049$. Expected hit rates were higher when they were asked about before, rather than after, interval size and error (6.4 vs. 6.0). There was also an interaction between Question Domain and Question Order, $F(1, 134) = 8.8$, $p = .004$. The participants reported lower hit rates for UCSD questions than

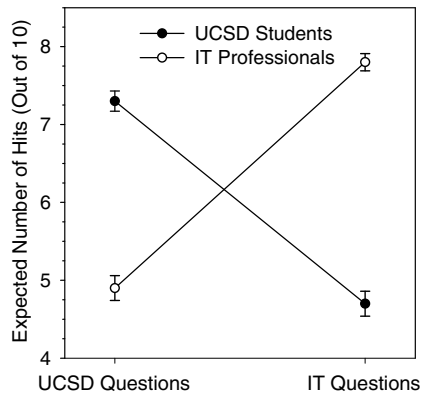


Fig. 3. Experiment 2 results for expected hit rate. Standard error bars are shown.

for IT questions when hit rates were asked about first (5.8 vs. 6.2), but this was reversed when hit rates were asked about last (6.4 vs. 6.3). This interaction accounts for the main effect of Question Order, although it is unclear why the interaction occurred.

Most important was the interaction, shown in Fig. 3, between Target Group and Question Domain, $F(1, 134) = 331.3$, $p < .001$. UCSD students were expected to have higher hit rates than IT professionals for the UCSD questions (7.3 vs. 4.7), but they were expected to have lower hit rates than IT professionals for the IT questions (4.9 vs. 7.8). Thus, participants expected experts to have higher hit rates than novices, even though this was not the case in Experiment 1 for these bounded questions (compare Figs. 1a and 3).

Expected interval size

A 2 (Question Domain) \times 2 (Question Order) mixed-model ANOVA on expected interval size revealed only a main effect of the within-subjects Question Domain variable, $F(1, 134) = 172.3$, $p < .001$. The participants expected the IT professionals to report wider intervals than the UCSD students for the UCSD questions (1.1 on the -3 to 3 scale) and they expected the UCSD students to report wider intervals than the IT professionals for the IT questions (-1.5). These expectations are qualitatively accurate with respect to the results of Experiment 1. (Note that because the participants answered four questions for expected hit rate but only two questions for expected interval size [and error, below], the fact that participants expected better performance for experts resulted in an interaction for hit rate but a main effect for interval size. Using the comparative rating scale eliminated “Target group” as a variable for the interval size and error questions.)

Expected interval error

A 2 (Question Domain) \times 2 (Question Order) mixed-model ANOVA on expected interval error revealed only

a main effect of the within-subjects Question Domain variable, $F(1, 134) = 371.7$, $p < .001$. Participants expected the UCSD students’ intervals to have less error than the IT professionals’ for the UCSD questions (1.6 on the -3 to 3 scale) and they expected the IT professionals to have less error than the UCSD students for the IT questions (-1.6). This is again qualitatively accurate with the results of Experiment 1.

Discussion

The question addressed in this experiment was whether people are aware of the advantages that experts do and do not have to offer, as shown in Experiment 1. With respect to hit rates, the participants expected experts to be much less overconfident than novices. This does not accurately describe performance in Experiment 1, where experts and novices were similarly overconfident in the bounded questions. Participants expected experts to have a hit rate of 75%, although their actual hit rate was only 44% in Experiment 1. Participants correctly predicted that experts would be overconfident, but they underestimated the degree of overconfidence. In this respect, our results suggest that people may consult experts for the wrong reason.

The fact that the participants—UCSD students—expected IT professionals asked about the IT industry to have a relatively high hit rate shows that people (incorrectly) expect experts to perform better than novices and to perform better than they actually do. In addition, the fact that the participants expected other UCSD students asked about UCSD to have a high hit rate suggests that experts expect *themselves* to perform better than novices and better than they actually do.

Although participants were not good at predicting experts’ hit rates, they predicted novice performance reasonably well. They expected novices’ 90% confidence intervals to have hit rates of 48%, which is roughly in accord with the typical empirical findings for 90% confidence intervals in general (approximately 50% hit rates), and with the results of Experiment 1 in particular (41% for the bounded questions).

Participants also did a good job of predicting the qualitative results for both interval size and interval error. In Experiment 1, experts reported narrower intervals with midpoints closer to the truth, and this is what participants in Experiment 2 expected.

Finally, it is worth noting that participants expected experts to report considerably narrower intervals than novices *and* to have much higher hit rates. Because narrower intervals imply a lower hit rate, this means one of two things. The first is that people simply expect experts to be better on all dimensions of performance without considering that some dimensions are in conflict. The second is that people know that narrower

intervals imply a lower hit in general, but they expect the increase in hit rate from experts' smaller interval errors to not just offset (as shown by the results of Experiment 1), but to overcome, the decrease in hit rate from narrower intervals.

General discussion

Experiment 1 showed that experts and novices were about equally overconfident when reporting 90% subjective confidence intervals. This was not, however, because experts and novices reported similar intervals. Instead, experts' intervals were narrower (more informative) and better centered on true values. Because narrower intervals decrease hit rate and intervals with less error increase hit rate, the net effect was that overconfidence remained largely the same. This explains why experts and novices tend to be about equally overconfident. So the bad news is that experts' intervals are unlikely to be well calibrated, or even better calibrated than novices. But the good news is that experts' interval estimates have less error and are more informative.

It is not surprising that experts' intervals have less error. Experts report more accurate point estimates (Önkal et al., 2003) and, if they center their intervals on these point estimates, lower error will result. It is less obvious, however, that experts would report narrower intervals, especially since novices tend to be so overconfident and a simple way to improve calibration would be to report wider intervals. One could imagine that experts have greater metaknowledge, or more knowledge about the limits of their knowledge, and widen their intervals accordingly, thereby improving hit rate and reducing overconfidence. Instead, we found that experts tended to report narrower intervals. This is consistent with results for novices by Yaniv and Foster (1997), who found that less error, which presumably indicated greater domain knowledge, was correlated with narrower intervals (but see Önkal et al., 2003). Not only do experts report tighter intervals, they tighten their intervals in a manner that tends to offset the increased hit rate from reduced error. This is a very interesting phenomenon that deserves further study (see also Yaniv & Foster, 1997).

Thus, although a straightforward way to improve calibration is to report wider intervals, this seems to go against the deeply ingrained tendency—for both experts and novices—to balance error ($|t - m|$) and interval width (g). Yaniv and Foster's (1995, 1997) explanation of novices' narrow intervals (and resulting low hit rates) in general is that people care not only about hits, but about informativeness as well. Narrow intervals convey more information and people strive to report informative intervals. Yaniv and Foster also

pointed out an interesting asymmetry in feedback with respect to interval width and hit rate. Whereas an interval's informativeness can be assessed immediately, whether it contains the true value may not be known for a long time, if at all. Furthermore, one must assess many $X\%$ confidence intervals before knowing if the producer is well calibrated, making calibration very difficult to evaluate, both for the recipients of the intervals and for the producers themselves. Therefore, producers of intervals may, in accord with the feedback they receive, focus more on informativeness and less on hit rate.

While Experiment 1 revealed which dimensions of performance experts were good at, Experiment 2 examined which dimensions people expected experts to be good at. It was found that both experts and novices expected experts to be much better calibrated than novices and much better than they really were. Thus, experts may be consulted for the wrong reason, and even experts appear unaware of this. However, Experiment 2 also showed that both experts and novices correctly expected experts to provide narrower intervals that were better centered on the truth. Anyone considering consulting experts for interval estimates should know the benefits and limitations of expertise in this context. In particular, acting on the assumption that experts' intervals have a high probability of containing the true value would be a mistake. Moreover, it is worth noting that the asymmetry in feedback we noted above is consistent with the results of Experiment 2: participants correctly expected experts to provide relatively informative intervals, but they incorrectly expected experts to be much better calibrated. It might be relatively easy to learn about informativeness, but difficult to learn about calibration.

In their review of expert judgment, Camerer and Johnson (1997) addressed the “process-performance paradox”: How can experts know so much and predict so poorly? They concluded that prediction is but one task that experts perform, and that “...experts do some things well and others [e.g., prediction] poorly” (p. 357). A well known example of this general conclusion is that experts are indispensable for measuring variables (Sawyer, 1966) and discovering new ones (Johnson, 1988), but they are poor at combining diverse sources of information in order to arrive at a single predictive judgment (e.g., Dawes, Faust, & Meehl, 1989). Our results can be viewed similarly. Experts are good at reporting relatively narrow intervals centered on true values, but they are no better than novices at reporting well calibrated, high-confidence intervals. Making good decisions about whether to consult experts in this context requires understanding what experts are, and are not, likely to deliver.

Appendix A

The 40 questions used in the survey (and their true values). Though the questions are grouped here, there were two random orders in the experiment.

UCSD bounded (percentage) questions

1. In 2001, what percent of UCSD students were graduate students? (13.9%)
2. In 2001, what percent of UCSD freshmen did not return for their sophomore year? (7%)
3. In 2001, what percent of UCSD freshmen had a college GPA of at least 2.5? (65.1%)
4. In 2001, what percent of UCSD undergraduates reported being from minority ethnic groups? (48.2%)
5. In 2001, what percent of applying freshmen did UCSD admit? (43%)
6. In 2001, what percent of UCSD undergraduates were enrolled as Biology majors? (17.4%)
7. In 2001, what percent of entering freshmen had a high school grade point average of at least 3.9? (62%)
8. In 2001, what percent of UCSD undergraduates received some type of financial assistance (includes loans, grants, work-study, and scholarships)? (59%)
9. In 2001, what percent of UCSD freshmen lived off campus? (20%)
10. In 2001, what percent of UCSD students were from San Diego? (24%)

IT bounded (percentage) questions

1. As of January 2001, what percent of Americans used online banking services? (5%)
2. As of September 2002, what percent of Americans went online every day? (35%)
3. As of September 2002, broadband connectivity was available to what percent of US households? (80%)
4. As of February 2002, what percent of American households were connected to the internet? (60%)
5. As of September 2002, what percent of the operating system market did Apple hold (worldwide)? (1.43%)
6. What was Microsoft's gross profit as a percent of revenue in 2001? (29%)
7. What was the percent decrease in the NASDAQ (stock exchange) between January 1, 2002 and October 1, 2002? (37.8%)
8. As of September 2002, what percent of people (worldwide) using search engines used Google? (55.1%)
9. What was the US unemployment rate (%) in September 2002? (5.6%)

10. As of September 2002, what percent of the US population had ever been online? (67%)

UCSD unbounded questions

1. Where is UCSD ranked on US News and World Report's list of best universities (public only—2001)? (7th)
2. For those students who became employed, what was the average starting salary of a graduating UCSD undergraduate (2001)? (\$39,700)
3. As of October 2001, how many full-time professors were employed by UCSD? (820)
4. How many UCSD undergraduates were enrolled as Economics majors in 2001? (1456)
5. In 2001, what was the average SAT I score of an entering UCSD freshman (math and verbal combined)? (1264)
6. How many freshmen applications did UCSD receive for the 2001–2002 school year? (38,187)
7. What is the distance from UCSD campus to downtown San Diego (miles)? (12)
8. How many UCSD undergraduates were female in 2001? (9105)
9. How many undergraduates were enrolled at UCSD in 2001? (17,505)
10. As a California resident, what is the estimated cost of attending UCSD as an undergraduate, for the 2002–2003 school year (includes tuition, books, room, board, and other expenses)? (\$15,975)

IT unbounded questions

1. How much did the average System's Analyst living in San Francisco, CA earn in 2001? (\$61,458)
2. As of March 2002, how many people were employed in Information Technology (US)? (9.98 million)
3. In what year did Microsoft release its first version of Microsoft-Windows? (1985)
4. How much money did US businesses spend on software in 2001? (\$189 billion)
5. As of June 2002, how many people subscribed to AOL's ISP (Internet Service Provider) service? (35.1 million)
6. What were IBM's total revenues in 2001? (\$85.87 billion)
7. How many Personal Computers (PCs) were shipped in 2001 (worldwide)? (134 million)
8. What is the total maximum dollar amount available to an individual under California unemployment? (\$17,160)
9. How many people are currently employed by Hewlett-Packard (HP)? (145,000)
10. How many companies in the Fortune 500 had revenues above \$5 billion in 2002? (341)

Appendix B

We conducted an experiment during the fall of 2002 and we want you to try to predict the results.

We asked both UCSD students and information technology (IT) professionals questions about both UCSD and the IT industry. The UCSD students were participating in an experiment for credit (just like you) and most of the IT professionals were employed by leading companies like IBM and Hewlett-Packard. No one was paid for participation.

Everyone was asked to provide *interval estimates* of values they were presumably uncertain of. In particular, they were asked to provide a low value and a high value such that they were *90% confident* that the resulting interval contained the true value. That is, they were to expect that 9 out of 10 of their reported intervals would contain the true value.

As one example, all participants were asked the following UCSD question:

	90% Confidence Interval	
	Low	High
In 2001, what percent of UCSD students were graduate students?		

Participants then reported a low value and a high value such that they were 90% confident the true value fell inside the resulting interval. Note that the above question asks for a percentage, so the reported values were between 0 and 100. In fact, all the questions asked for percentages, so all responses were between 0 and 100.

On the following page are 10 UCSD questions and 10 IT questions that both UCSD students and IT professionals answered. We *do not* want you to report any intervals (which is why we put X's in the spaces for responses). Just look over the questions so you know exactly what the participants' task was. Remember that all participants answered all questions.

If you ever have any questions, please ask the experimenter. After you have read and understood the above instructions, please turn the page.

References

Alpert, M., & Raiffa, H. (1982). A progress report on the training of probability assessors. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 294–305). New York: Cambridge University Press.

Camerer, C. F., & Johnson, E. J. (1997). The process-performance paradox in expert judgment: How can experts know so much and

predict so badly? In W. M. Goldstein & R. M. Hogarth (Eds.), *Research on judgment and decision making: Currents, connections, and controversies* (pp. 342–364). Cambridge: Cambridge University Press.

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

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical vs. actuarial judgment. *Science*, 243, 1668–1674.

Johnson, E. (1988). Expertise and decision under uncertainty: Performance and process. In M. T. H. Chi, R. Glaser, & M. J. Farr (Eds.), *The nature of expertise* (pp. 209–228). Hillsdale, NJ: Erlbaum.

Juslin, P., Wennerholm, P., & Olsson, H. (1999). Format dependence in subjective probability calibration. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25, 1038–1052.

Keren, G. (1987). Facing uncertainty in the game of bridge: A calibration study. *Organizational Behavior and Human Decision Processes*, 39, 98–114.

Klayman, J., Soll, J. B., González-Vallejo, C., & Barlas, S. (1999). Overconfidence: It depends on how, what, and whom you ask. *Organizational Behavior and Human Decision Processes*, 79, 216–247.

Koehler, D. J., Brenner, L., & Griffin, D. (2002). The calibration of expert judgment: Heuristics and biases beyond the laboratory. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 686–715). Cambridge: Cambridge University Press.

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance*, 20, 159–183.

Lichtenstein, S., Fischhoff, B., & Phillips, P. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 306–334). New York: Cambridge University Press.

McKenzie, C. R. M., & Amin, M. B. (2002). When wrong predictions provide more support than right ones. *Psychonomic Bulletin and Review*, 9, 821–828.

Murphy, A. H., & Winkler, R. L. (1977). Reliability of subjective probability forecasts of precipitation and temperature. *Applied Statistics*, 26, 41–47.

Önköl, D., Yates, J. F., Simga-Mugan, C., & Öztin, Ş. (2003). Professional vs. amateur judgment accuracy: The case of foreign exchange rates. *Organizational Behavior and Human Decision Processes*, 91, 169–185.

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

Russo, J. E., & Schoemaker, P. J. H. (1992). Managing overconfidence. *Sloan Management Review*, 33, 7–17.

Sawyer, J. (1966). Measurement and prediction, clinical and statistical. *Psychological Bulletin*, 66, 178–200.

Snizek, J. A., & Buckley, T. (1991). Confidence depends on level of aggregation. *Journal of Behavioral Decision Making*, 4, 263–272.

Soll, J. B., & Klayman, J. (2004). Overconfidence in interval estimates. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30, 299–314.

Stael von Holstein, C.-A. S. (1972). Probabilistic forecasting: An experiment related to the stock market. *Organizational Behavior and Human Performance*, 8, 139–158.

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

Tomassini, L. A., Solomon, I., Romney, M. B., & Krogstad, J. L. (1982). Calibration of auditors' probabilistic judgments: Some

- empirical evidence. *Organizational Behavior and Human Performance*, 30, 391–406.
- Yaniv, I., & Foster, D. P. (1995). Graininess of judgment under uncertainty: An accuracy-informativeness trade-off. *Journal of Experimental Psychology: General*, 124, 424–432.
- Yaniv, I., & Foster, D. P. (1997). Precision and accuracy of judgmental estimation. *Journal of Behavioral Decision Making*, 10, 21–32.
- Yates, J. F., McDaniel, L. S., & Brown, E. S. (1991). Probabilistic forecasts of stock prices and earnings: The hazards of nascent experience. *Organizational Behavior and Human Decision Processes*, 49, 60–79.