Probabilities associated with precise and vague forecasts

Erik Løhre¹ and Karl Halvor Teigen^{1,2}

¹Simula Research Laboratory, Oslo

² Department of Psychology, University of Oslo

Corresponding author:

Erik Løhre

Simula Research Laboratory

P.O. Box 134, 1325 Lysaker, Norway

Telephone: +47 41 42 01 35

E-mail: erikloh@simula.no

The published version of this paper can be found in the Journal of Behavioral Decision Making, doi: 10.1002/bdm.2021

Word count: 10483 words (including abstract, references, and tables)

Probabilities associated with precise and vague forecasts

Abstract (185 words)

Forecasts of future outcomes, such as the consequences of climate change, are given with different degrees of precision. Logically, more precise forecasts (e.g., a temperature increase of 3-4 degrees) have a smaller probability of capturing the actual outcome than less precise forecasts (e.g., a temperature increase of 2-6 degrees). Nevertheless, people often trust precise forecasts more than vague forecasts, perhaps because precision is associated with knowledge and expertise. In five experiments, we ask whether people expect highly confident forecasts to be associated with wider or narrower outcome ranges than less confident forecasts (Experiments 1, 2 and 5), and, conversely, whether they expect precise forecasts to be issued with higher or lower confidence than vague forecasts (Experiments 3 and 4). The results revealed two distinct ways of thinking about confidence intervals, labeled distributional (logical) and associative. Distributional responses occurred somewhat more often in within-subjects designs, where wide and narrow prediction intervals and high and low probability estimates can be directly compared, whereas separate evaluations (in between-subjects design) suggested associative responses to be slightly more frequent. These findings are relevant for experts communicating forecasts through confidence intervals.

Keywords: precision, informativeness, confidence, intervals, uncertainty

Forecasts of future outcomes, such as the consequences of climate change, are given with different degrees of precision. For instance, a projected temperature increase of 0.4 to 0.6 degrees gives a more precise estimate about what the outcome might be than a temperature increase of 0.1 to 0.9 degrees. Will the more precise estimate also be perceived as more, or less certain? On one hand, a narrow (precise) interval seems to indicate more certainty about the range of potential outcome values. At the same time, narrow intervals are quite *uncertain* in the sense that they have a smaller probability of capturing the actual outcome than a wide interval. Thus, logically, a forecaster may have a choice between predicting a narrow range of outcomes with a low probability, or a more probable prediction of a wider potential range. Recipients would prefer predictions to be both exact and highly probable, but in areas fraught with uncertainty forecasters have to strike a balance between the two, such as when the Intergovernmental Panel on Climate Change (IPCC) asserts that the “global mean surface temperature change for the period 2016-2035 relative to 1986-2005 will *likely* be in the range of 0.3°C to 0.7°C” (IPCC, 2013, p. 20). Here the predicted outcome is given as an interval, which is claimed to be *likely* ($p > .66$, according to the guidelines), rather than *very likely* ($p > .90$), or *virtually certain* ($p > .99$), which would require a wider range of temperatures.

We examine in the present paper lay people’s intuitions about the trade-off between probabilities and prediction intervals in a context of climate change forecasts. In contrast to many previous studies, where people provide their own intervals for a specified confidence level, our participants were asked which interval (or confidence level) another forecaster might have in mind. Specifically, our research question is whether people expect a high-probability forecast to be associated with a wider or a narrower outcome range than a low-probability forecast (Experiments 1, 2, and 5), and, conversely, whether they expect a narrow interval forecast to be issued with a higher or lower probability than a wider interval (Experiments 3 and 4). Formally, one might expect a negative relationship between preciseness and probabilities, reflecting the compensatory nature of probabilities and prediction intervals. Responses in accordance to this principle will henceforth be described as *distributional*. However, since probabilities and precision both can indicate degree of certainty/uncertainty, some people may find it natural to imagine a positive relationship between the two, expecting high preciseness to go with high probabilities, perhaps because both appear to reflect the forecaster’s level of expertise. Responses indicating a positive link will henceforth be described as *associative*. We do not claim that these labels accurately describe the cognitive mechanisms involved, but believe that “associative” responses are more spontaneous and immediate, and require less deliberation than “distributional”

responses, and might be typical of situations where predictions cannot be compared, but are evaluated in a separate mode (Hsee, 1996).

Overprecision and the preciseness paradox

The relationship between probabilities and interval estimates has been extensively studied and discussed in the literature about confidence intervals. People have been regularly shown to produce confidence intervals that are far too narrow, or *overprecise*, to use Moore and Healy's (2008) apt term. Overprecision has typically been studied by asking people to produce intervals corresponding to an assigned, high level of confidence (90% or more). However, people seem rather insensitive to the relationship between probability and interval width. In within-subjects designs where the same participants are asked to create 90% and 50% intervals for the same question, they seem to realize that the 90% interval must be wider, although not to the extent dictated by a normal probability distribution (Budescu & Du, 2007). When different groups are instructed to estimate 90% and 50% intervals, they do not appear to be sensitive to the assigned confidence level, and produce intervals of the same width (Teigen & Jørgensen, 2005; Langnickel & Zeisberger, 2016), leading to overprecise 90% intervals but more well-calibrated 50% intervals. Thus overprecision can in part be attributed to a deficient understanding of probability distributions (Moore, Tenney, & Haran, 2016).

Conversational norms, which ask the speaker to "make your contribution as informative as required" (Grice, 1975, p. 45), may also play a role. Yaniv and Foster (1995) found that people often preferred an informative (narrow) estimate to an imprecise (wide) estimate, even when the wide estimate was technically correct and the narrow estimate was not. Similarly, more recent studies show that narrow range forecasts were preferred over wide range forecasts in the financial domain (Du, Budescu, Shelly, & Omer, 2011), and for software project effort estimates, a developer giving a narrow interval estimate was rated as more competent and trustworthy than a developer giving a wide interval estimate (Jørgensen, 2016).

Judgment producers are also aware of the preference for informative statements, and try to make their estimates as precise as their knowledge allows them to be (Yaniv & Foster, 1997). This enables recipients of the communication to draw inferences from interval width to level of expertise (McKenzie, Liersch, & Yaniv, 2008), something strategic forecasters can exploit by using forecast precision to convey authority when they need to appear in control (Hayward & Fitza, 2016). But by producing tighter intervals, experts remain overprecise even within their domain of expertise (Connolly & Dean, 1997). Thus, both speakers and listeners

seem to disagree with logician Carveth Read's statement (1920) that "it is better to be vaguely right than exactly wrong".

A precise statement can be taken as a sign of expertise and thus appear more convincing than an imprecise one, even when the latter is more probably true. Teigen (1990) investigated this "preciseness paradox" by asking participants which statement in a set they would feel most confident about. Overall, people indicated they would be more confident about the most precise statement, compared to a more vague statement that included the precise one and thus had a better chance of capturing the true value. But precise statements are not always trusted. In Teigen's study, participants were less convinced about precise statements about the future than about the past. Moreover, when asked which statement they would be most *skeptical* about, they again selected the most precise statement. This shows an inherent ambiguity in precise statements: they are desired for being informative and are associated with high confidence and expertise, yet people are able to realize (when given the proper frame) that they are also more risky or *bolder*. Thus, when a precise statement is correct it is more impressive than when a vague statement is correct, and bold predictions that are slightly incorrect may even be preferred to correct, but timid predictions (McKenzie & Amin, 2002).

The power of precision has also been demonstrated in several recent studies comparing precise numbers (e.g., 21 or 19) to nearly equivalent round numbers (e.g., 20). People believe that speakers using precise numbers are more confident than speakers using round numbers (Jerez-Fernandez, Angulo, & Oppenheimer, 2014), and precise numbers were found to act as stronger anchors than round numbers both in price negotiation contexts (Loschelder, Stuppi, & Trötschel, 2014; Mason, Lee, Wiley, & Ames, 2013) and in a more traditional anchoring paradigm (Janiszewski & Uy, 2008; Zhang & Schwarz, 2013). But again, the precision advantage is not universal: no effect of anchor precision was found for estimates of software projects (Løhre & Jørgensen, 2016), perhaps because such projects are not believed to be completely predictable in advance. Du et al. (2011) have suggested that people expect different degrees of (im)precision in different domains. Most people expect precision in some contexts (e.g., medicine), but not in others (e.g., political forecasting).

Thus, in studies using several different approaches, we find evidence for two different ways of thinking about precision. People will often assume precision indicates knowledge and expertise, and expect precision and high probabilities to go together (be associated), but will under some circumstances realize that precise claims are risky and should be met with

skepticism, indicating a negative relationship between precision and confidence more in line with distributional logic.

Probabilities and levels of confidence

The certainty and uncertainty of a prediction is reflected not only by the range of potential outcomes, but also by the probabilities attached to outcomes in this range. High probabilities indicate a high degree of certainty, and are preferred to low probabilities for similar reasons as precise estimates are preferred to vague ones. For instance, high probability estimates are regarded by lay people both as more informative and more useful than low probabilities (Keren & Teigen, 2001b), and listeners assume that forecasters giving high probability predictions have conducted more in-depth analyses and can be trusted more than forecasters giving low probability predictions (Bagchi & Ince, 2016). In the courtroom, testimony from confident witnesses is given more weight than testimony from uncertain witnesses (Brewer & Burke, 2002). Confident advisors are believed to be more knowledgeable and hence more accurate than those who express less certainty, even when the latter are better calibrated (Yates, Price, Lee, & Ramirez, 1996). In other words, people make use of a “confidence heuristic” (Price & Stone, 2004), and rely more on confident than on less confident advisors (Snizek & Van Swol, 2001) – especially when information about hit rates is unavailable or costly to obtain (Sah, Moore, & MacCoun, 2013; Tenney, MacCoun, Spellman, & Hastie, 2007). However, studies indicate that degree of confidence (subjective probability estimates) is only weakly related to actual hit rates in several domains. Thus, a meta-analysis concludes that “confidence is apparently not a good proxy to accuracy” (Miller, Spengler & Spengler, 2015, p. 12).

Certainty expressed in terms of probabilities may be of two kinds: *internal*, reflecting an individual’s degrees of belief or knowledge, or *external*, reflecting randomness, frequencies and propensities in the external world. Confidence estimates of one’s own answer to trivia questions, like the height of the Eiffel tower, are clearly of the internal kind (since the actual height of the Eiffel tower is not subject to variation). In contrast, uncertainty about the next flip of a coin is usually attributed to random processes in the external world and not to the observer’s lack of knowledge about coin tosses. Complex forecasts may have an element of both: forecasts of climate change are perceived to be uncertain partly because of the complexity and unpredictability of the climate system itself, and partly due to the forecaster’s limited knowledge of processes that govern climate changes.

The distinction between internal (epistemic) and external (aleatory) uncertainty is not only of interest to probability theorists (Gillies, 2000; Hacking, 1975), but may also be reflected in lay people's thinking and in the language they use (Kahneman & Tversky, 1982). Some linguistic terms (like "a chance") appear more related to external probabilities whereas other terms (like "confidence") point to degrees of knowledge and belief (Fox & Ülkümen, 2017; Teigen & Løhre, 2017; Ülkümen, Fox, & Malle, 2016). External probabilities can be conceived as objective and calculable, taking an impersonal pronoun ("It is 80% probable") whereas internal probabilities rely on a subjective, or personal judgment ("I am 90% certain"), and direct attention towards the speaker and his or her personal opinions. *Certain* is an amphibian term: It has been identified as primarily epistemic, reflecting the speaker's belief or knowledge (Ülkümen et al., 2016), but the term can also be used in an external, more objective sense in phrases with impersonal pronouns. In a recent study (Løhre & Teigen, 2016) comparing externally and internally focused expressions we found that speakers were expected to express a higher degree of personal than impersonal certainty ("It is 70% certain" corresponded roughly to "I am 80% certain" in a between-subjects design).

It is not obvious that internal and external probabilities will be associated with interval estimates in the same way. Knowledgeable speakers are generally expected to be confident and to express themselves with a high degree of precision. Therefore, people might believe that precise statements (narrow intervals) are accompanied by higher personal confidence estimates than less precise statements, in accordance with the *associative* view. External probabilities, on the other hand, might suggest calculations, and induce a more formal, mathematical mindset. This in turn indicates a negative rather than positive association between preciseness and probability, in line with the *distributional* view. Hence, probabilities will in all the following studies be presented in two formats: impersonal and external ("there is a 90% probability") and personal and internal ("I am 90% certain").

Experiment 1

In this experiment we asked participants to suggest upper and lower interval limits of forecasts coming from two different experts, who expressed their estimates of the same events with either high (90%) or low (60%) probability. In one condition, probabilities were expressed in an impersonal format ("there is a X% probability") and in another as personal confidence ("I am X% certain"). The direct comparison of forecasters was believed to make the logical issue transparent and facilitate formal, distributional thinking. This would allow

participants to realize that 90% intervals should be wider than those that are merely 60% certain.

Method

Participants. Psychology students at the University of Oslo were invited to complete questionnaires in a break between lectures. Three questionnaires were discarded due to extreme responses and obvious misunderstandings (e.g., minimum value higher than the maximum value). Of the 90 remaining participants, 69 were women and 20 were men (one did not indicate gender), mean age 24.3 years ($SD = 6.4$, two participants did not indicate age).

Questionnaires. The participants were told to imagine that a journalist had asked two leading climate scientists to give their forecasts for global temperature and sea level, as well as the precipitation in Norway, by the year 2050. Participants were asked to complete the scientists' probability statements with an interval that appeared reasonable. A translation of the first scenario is given below.

“There are several different models for how the global average temperature will be in 2050. On average, the models show that the temperature will increase by about 1.5°C, but there is significant variation between models, depending upon their assumptions. Evensen says: There is a 90% probability [I am 90% certain] that the global mean temperature will increase by between _____ and _____ °C within the year 2050. (Write numbers in both blank spaces) Martinsen says: There is a 60% probability [I am 60% certain] that the global mean temperature will increase by between _____ and _____ °C within the year 2050. (Write numbers in both blank spaces)”

For all three scenarios and in both conditions, Evensen expressed a higher level of probability/certainty (90%) than Martinsen (60%), the order of experts being counterbalanced across subjects. All scenarios included a mean value (1.5°C in the example above) as an anchor point to minimize individual differences in participants' estimates due to personal opinions of climate change. Probability format (internal: I am X% certain, external: there is a X% probability) was different in the two conditions. We will henceforth refer to both these formats as “levels of confidence”.

After completing the intervals, participants rated perceived agreement between the two experts on a scale from 1 (They seem to disagree completely) to 7 (They seem to agree completely), and indicated which scientist appeared to be a greater expert on climate change, along with a short explanation for their choice.

Results

Participants were split in their views on the relationship between probability/certainty and width of prediction interval. It is apparent from Figure 1 that the number of participants who thought that a 90% level of confidence warranted a wider prediction interval than a 60% level of confidence (as required by a distributional model) were not in majority, as an approximately equal number thought that a higher confidence level would be associated with a more precise forecast. An additional, slightly smaller number, suggested that both intervals were equal. The proportion of distributional responses was somewhat, but not significantly larger in the external than in the internal condition (44.4% vs. 37.0%, respectively). Differences were tested by separate 2 (probability format: internal vs. external) x 2 (confidence level: 60% vs. 90%) mixed ANOVAs on the interval widths for each scenario, which showed no significant effects of either factor nor any interaction, all F 's < 2.7 , p 's $> .11$.

<Insert Figure 1 about here>

In a normal probability distribution 90% of the area lies within $\pm 1.64 z$ and 60% within $\pm 0.84 z$. Thus if outcome estimates follow a normal distribution, 90% uncertainty intervals should be nearly twice as large as intervals that are merely 60% probable (assuming that the forecasters are of equal expertise). And among the subset of respondents who believed that 90% intervals were larger, the median answer was in fact that 90% intervals were two times larger than 60% intervals.

Experts were rated as being in relatively good agreement with each other (mean ratings between 4.33 and 4.92 in the different conditions), with no significant effects of the condition, order, nor any interaction (all p 's $> .14$). Of the 83 participants who answered the question of expertise, about half indicated a preference for the less confident expert. This preference was slightly, but not significantly, stronger in the external than in the internal condition (65% vs. 45%), $\chi^2(1, N=83) = .885$, $p = .347$. Some participants argued that 90% is “too certain” given the uncertainty of the subject, while other participants argued that a higher

confidence level appeared more credible. A few participants also mentioned the precision of the estimate, often claiming that a narrow interval appears more accurate.

The main conclusion to be drawn from Experiment 1 is that people do not consistently expect 90% intervals to be larger than 60% intervals, even under conditions where both intervals are available for direct comparison. Instead, “distributional” and “associative” responses were about equally frequent.

Experiment 2

Experiment 2 was basically a between-subjects version of the first experiment. In this design, participants did not have the opportunity to consider the implications of two levels of probability/certainty presented side by side, and might rely more on their immediate intuitions than on formal deliberations. We hypothesized that people intuitively see low confidence as a signal of uncertainty, suggesting wide intervals, while high confidence indicates expertise and knowledge, including a high degree of precision. We expected accordingly by this design more associative than distributional responses. Thus 90% confidence levels would be completed with narrower, not wider, intervals than 60% confidence levels.

Method

Participants. MTurk workers from the U.S. were invited to participate in a short experiment for a payment of \$0.75. We excluded 30 people who failed a simple attention check, or who gave extreme or erroneous responses (e.g., higher minimum than maximum values). The final dataset included 242 participants with a mean age of 33.7 years ($SD = 10.4$, two participants did not indicate age), 59.5% were male, and 88% reported to have at least some college education, with “College graduate (4 years)” being the most frequent response (43% of the participants). Participants were randomly assigned to one of four conditions in a 2 x 2 between-subjects design, with confidence level (90% vs. 60%) and probability format (internal vs. external) as the two factors.

Materials and procedure. The questionnaires described a climate scientist, Dr. Johnston, who is asked how climate change will affect different (fictional) species around the U.S. The participants were asked to complete three forecasts that Dr. Johnston had made, by filling in plausible minimum and maximum estimates. The order of scenarios was randomized across subjects. One scenario concerned the fine-spotted bull trout, and was phrased as follows:

“The fine-spotted bull trout population is predicted to decline as a consequence of climate change. On average, the different models show that the fine-spotted bull trout population will decline by about 32% within the year 2050, but there is significant variation between models, depending on which assumptions they are based on. Dr. Johnston says: “Within the year 2050, there is a 90% (60%) probability [I am 90% (60%) certain] that the fine-spotted bull trout population will decline by minimum _____ % and maximum _____ %.”

The two other scenarios described a predicted increase in the area suitable for the yellow spring moth in Illinois and a predicted northwards expansion of the territory of the California silver fir. Dr. Johnston’s statements indicated either a high (90%) or a low (60%) level of confidence, expressed in an internal (I am X% certain) or an external format (There is a X% probability).

At the end, participants were asked to assume that Dr. Johnston had answered the newspaper’s questions as they had indicated, and to rate their agreement with the following statements (on a scale from 1: completely disagree, to 7: completely agree): “Dr. Johnston seems like a competent scientist”, “Dr. Johnston’s predictions seem trustworthy”, “There seems to be a great deal of uncertainty about the consequences of climate change for different species”, and “Dr. Johnston’s statements would make me feel worried about the consequences of climate change”.

Results

<Insert Table 1 about here>

To analyze the effects of the independent variables across scenarios, we computed the width of the participants’ intervals for each scenario (see Table 1 for descriptives). These were then log-transformed due to skewness, z-transformed and averaged. A 2 x 2 ANOVA of these averaged intervals showed no effect of probability format and no interaction (p ’s > .36), but a significant effect of confidence level, $F(1,238) = 6.143$, $p = .014$, $\eta^2_p = .025$, with the intervals in the low confidence conditions being wider ($M = .14$, $SD = .78$) than the intervals in the high confidence conditions ($M = -.13$, $SD = .91$). Further inspection of the distribution of responses showed that 68.1% of the intervals in the low confidence conditions were wider than the mean interval size in the high confidence conditions, and 60.3% of the intervals in the high confidence conditions were narrower than the mean interval size in the low

confidence conditions. Thus participants gave more associative than distributional responses, in line with the hypothesis that high confidence is intuitively associated with precision.

Separate ANOVAs for the minimum and maximum values showed no difference in maximum values in the different conditions, all F 's < 1.1 , while minimum values were significantly lower in the low confidence condition ($M = -.13$, $SD = .83$) than in the high confidence condition ($M = .12$, $SD = .78$), $F(1,238) = 5.938$, $p = .016$, $\eta^2_p = .024$ (there was no effect of probability format nor any interaction, F 's < 1).

Interval width (as suggested by participants) correlated negatively with ratings of competence ($r = -.14$, $p = .026$) and trust ($r = -.21$, $p = .001$), and positively with ratings of uncertainty ($r = .17$, $p = .008$), but was not significantly related to ratings of personal worry ($r = -.10$, $p = .107$). Thus experts with wide intervals were perceived to be more uncertain and less competent and trustworthy than experts with narrower intervals.

Ratings were also affected by probability format (confidence level had no significant effects, nor were there any interactions, p 's $> .16$). When Dr. Johnston used an external phrase (there is a 90% probability), he was considered to be somewhat more competent and trustworthy than when he used a corresponding internal phrase (I am 90% certain), $F(1, 238) = 4.596$, $p = .033$, $\eta^2_p = .019$, and $F(1, 238) = 4.752$, $p = .030$, $\eta^2_p = .020$, for competence and trustworthiness, respectively. Internal phrases were also judged to reflect significantly more uncertainty, $F(1, 238) = 10.546$, $p = .001$, $\eta^2_p = .043$. Ratings of worry were not affected by the experimental factors, all p 's $> .22$.

Discussion

While Experiment 1 led to a similar number of distributional and associative responses, Experiment 2 suggested that people intuitively associate high confidence levels with narrow intervals in a between-subjects design. These two features may be expected to go together since they both express certainty and indicate knowledge and expertise.

Higher confidence levels indicated primarily higher minimum values in this experiment. This might point to an association between confidence level and the *strength* of the predicted outcome. In all the current scenarios, higher values meant a “stronger” outcome. For instance, if the trout population is expected to decline by *at least 20%*, it seems more certain that *something* is happening to the trout than if a decline of *at least 10%* is predicted. This association between confidence level and outcome strength is compatible with a propensity interpretation of probability, where higher probabilities are thought to signal stronger (and more immediate) outcomes (Keren & Teigen, 2001a; Løhre, 2017, for more on this topic).

Experiment 3

In both previous experiments, people were given pre-assigned confidence levels, and asked which intervals they expected from the experts. But the relationship between probability and interval width can also be approached from the opposite direction. Hence, in the next two experiments, participants were assigned to narrow and wide interval conditions, and asked to infer the expert's degree of confidence. The studies were designed as a within-subjects version (Experiment 3) and a between-subjects version (Experiment 4) of the same basic experiment and were run at the same time, with similar materials, and participants randomly allocated to Experiment 3 or 4. As distributional responses require an elementary understanding of formal rules, we included in these experiments a measure of statistical numeracy, expecting distributional responses to be more frequent among highly numerate participants.

Method

Participants. Participants were recruited through MTurk and received \$1.10 for completing a questionnaire that was estimated to take about 10-20 minutes. We excluded 20 people who failed a simple attention check or spent less than 3 minutes on the survey. This left 394 participants (44.2% females), with a mean age of 36.9 years ($SD = 11.9$, two did not indicate age). The majority (81.7%) reported having at least some college education, with the most frequent response (40.1%) being college graduate (4 years). There were 202 respondents in the within-subjects version (Experiment 3), and 192 respondents in the between-subjects version (Experiment 4).

Questionnaires. Participants were told that two climate scientists, Dr. Johnston and Dr. Stevens, had been asked to predict the consequences of climate change for different species (fish, insects, and plants) around the U.S. One scientist gave wide intervals, whereas the other's intervals in comparison were narrow (to ensure that the values appeared reasonable, both narrow and wide intervals were aligned with participants' responses in Experiment 2). The participants' task was to fill in what they felt would be appropriate confidence values, either in an external or an internal probability format. Thus, the trout scenario read as follows:

The fine-spotted bull trout population is predicted to decline as a consequence of climate change. On average, the different models show that the fine-spotted bull trout population will decline by about 31% within the year 2050, but there is significant

variation between models, depending on which assumptions they are based on. After inspecting the models, Dr. Johnston predicts that the population will decline by 25%-35%, while Dr. Stevens predicts that the population will decline by 15%-45%. What probabilities do you think the two scientists give for their predictions? [How certain do you think the two scientists are about their predictions?] Complete their quotes to the newspaper by filling in appropriate numbers in the statements below.

Dr. Johnston says: - Within the year 2050, there is a _____ % probability [I am _____ % certain] that the fine-spotted bull trout population will decline by minimum 25% and maximum 35%.

Dr. Stevens says: - Within the year 2050, there is a _____ % probability [I am _____ % certain] that the fine-spotted bull trout population will decline by minimum 15% and maximum 45%.

Participants received three scenarios, similar to those used in Experiment 2, in random order. The order of the scientists with narrow and wide intervals was counterbalanced across participants. After completing the quotes, they were asked to rate which scientist they thought was most competent and trustworthy, which scientist's statements indicated greater uncertainty about the consequences of climate change for different species, and which scientist's statements would make the participants personally feel most worried about the consequences of climate change. These ratings were done on scales from 1 to 7, with 1 meaning "definitely Dr. Johnston" (the narrow expert), 7 meaning "definitely Dr. Stevens" (the wide expert), and 4 indicating no difference. Finally, all participants were given a seven items numeracy test, consisting of the four item Berlin Numeracy Test (Cokely, Galesic, Schulz, Ghazal, & Garcia-Retamero, 2012) and three items from an earlier numeracy test (Schwartz, Woloshin, Black, & Welch, 1997). Since MTurk participants are known to have a lot of experience taking surveys (Paolacci & Chandler, 2014), and these two tests are often used to measure numeracy, participants were asked whether they had answered the numeracy questions before. A single item measuring personal belief in climate change was also included¹. After a simple attention check, participants were debriefed and thanked for their participation.

Results

¹ We do not report the results for this measure as personal belief in climate change did not relate to any of the dependent variables in Experiment 3.

Wide intervals were, on average, accompanied by slightly higher confidence level estimates than narrow intervals, as shown in Table 2 (the difference between wide and narrow intervals is significant, $F(1,198) = 10.777, p = .001, \eta^2_p = .052$). This means that participants overall completed the experts' statements in a "distributional" way. However, the difference in confidence between wide and narrow intervals was not large, considering the fact that the wide intervals were about three to four times wider than the narrow intervals. This should allow the experts in the wide conditions to be more than 90% confident (based on a normal distribution where narrow intervals are 60% certain, and assuming that the forecasters are of equal expertise). However, among participants giving distributional responses, wide intervals were on average rated to be only 13.7% more certain/probable than narrow intervals ($M_{\text{wide}} = 76.6, SD = 16.3$, vs. $M_{\text{narrow}} = 63.0, SD = 10.5$).

<Insert Table 2 about here>

Table 2 further indicates that confidence level ratings were about 10 percentage points higher in the internal condition ($M = 74.2, SD = 15.7$) than in the external condition ($M = 65.0, SD = 18.1$), $F(1,198) = 14.466, p < .001, \eta^2_p = .068$, replicating previous findings (Løhre & Teigen, 2016). The table also suggests a somewhat stronger effect of interval width in the external probability format. However, the interaction of probability format and interval width did not reach significance, $F(1,198) = 3.034, p = .083, \eta^2_p = .015$.

To facilitate comparison between experiments, we also classified responses according to whether wide intervals received higher, lower or equal probability ratings compared to narrow intervals. Overall, a majority of 55.4% suggested that the expert with wide intervals would be more confident than the one with narrow intervals, against 31.7% of participants who indicated the opposite view, and 12.9% who thought they would be equally confident. As shown in Figure 2, distributional responses were more common in the external condition, but the difference between conditions was not significant, $\chi^2(2, N = 202) = 4.09, p = .129$.

<Insert Figure 2 about here>

Numeracy. About half of the participants (53.5%) indicated that they had answered all or some of the numeracy questions previously (i.e., in other experiments by different researchers). They scored higher ($M = 4.11, SD = 1.83$) than those who were new to the test

items ($M = 3.32$, $SD = 1.79$), $F(1,200) = 9.653$, $p = .002$, $\eta^2_p = .046$. For both groups, we found similar, positive correlations between numeracy and distributional responses. The overall correlation between numeracy and average difference scores (confidence estimates of wide intervals minus confidence estimates of narrow intervals) was $r = .29$, $p < .001$. This indicates a link between statistical numeracy and insight into the formal relationship between precision and confidence.

Ratings. The two experts were rated as approximately equally competent and trustworthy, and their statements were felt to cause an approximately equal amount of worry (mean ratings all between 3.6 and 4.5). Ratings of uncertainty indicated that the wide expert was generally seen as slightly more uncertain (all means > 4.3 , the overall mean was significantly different from 4.0, $t(201) = 6.553$, $p < .001$). There were no significant effects of probability format on any of the ratings, all p 's $> .092$. The average difference score for confidence estimates (wide interval minus narrow interval) correlated positively with ratings of competence ($r = .20$, $p = .004$) and trust ($r = .22$, $p = .001$), but not with ratings of uncertainty ($r = .01$, $p = .858$), or worry ($r = -.13$, $p = .062$). This indicates that participants with more distributional responses rated the wide expert as more competent and trustworthy and the narrow expert as causing slightly more worry.

Experiment 4

Method

This experiment was a between-subjects version of Experiment 3, with participants ($n = 192$) in separate conditions judging one scientist at the time rather than comparing the two. Thus Dr. Johnston predicted either that the fine-spotted bull trout population would decline by 25%-35% (narrow condition), or that it would decline by 15%-45% (wide condition). Participants completed the forecasts by suggesting an appropriate confidence level, either in an internal format ("I am X% certain") or an external format ("There is a X% probability"), yielding a 2 x 2 design with interval width and probability format as the two factors.

Results

<Insert Table 3 about here>

Mean confidence level estimates for all scenarios in four conditions are displayed in Table 3. A 2 x 2 ANOVA of mean estimated confidence reveals no main effect of probability format, $F(1,188) = 2.035$, $p = .155$, $\eta^2_p = .011$, or of interval width, $F < 1$, but a significant

interaction between the two factors, $F(1,188) = 5.425, p = .021, \eta^2_p = .028$.² Wide intervals were associated with slightly, but not significantly higher probabilities in the external probability format, $t(93) = -1.422, p = .158$, while the reverse relationship seems to hold for the internal format, $t(95) = 1.932, p = .056$. For comparison with previous experiments, we also inspected the distribution of responses, and found a majority of “distributional” responses in the external condition, with 59.2% of the narrow intervals receiving lower probabilities than the mean probability in the wide condition, and 63.0% of wide intervals given higher probabilities than the mean probability of narrow intervals. In the internal condition, “associative” responses were more frequent: 67.3% of the narrow intervals received higher probabilities than the mean probability in the wide condition, and 58.3% of wide intervals were given lower probabilities than the mean probability of narrow intervals. These effects suggest that people expected narrow intervals to go together with slightly higher personal certainty, which is consistent with the “associative” view, whereas impersonal probability estimates to some extent facilitated more distributional reasoning.

Numeracy. As in Experiment 3, about half of the participants (49%) indicated that they had answered the numeracy questions previously, and had a higher average numeracy score ($M = 4.15, SD = 1.97$) than those reporting no prior experience ($M = 3.37, SD = 1.83$), $F(1,190) = 8.133, p = .005, \eta^2_p = .041$. Numeracy scores correlated positively with estimated confidence level in the wide conditions ($r = .55, p < .001$, in the external wide condition, $r = .56, p < .001$, in the internal wide condition), but correlations were close to zero in the two narrow conditions.

To illustrate this effect, we used a median split to create groups of high and low numeracy participants. Low and high numeracy participants assigned similar confidence levels to narrow intervals, but suggested very different confidence levels for wide intervals, as seen in Figure 3. Participants with low numeracy thought wide intervals suggested low confidence, whereas more numerate participants thought wide intervals suggested high confidence. The fact that more numerate participants were willing to give higher probability/certainty ratings to wide intervals, again suggests that an understanding of statistics facilitates distributional thinking.

<Insert Figure 3 about here>

² Personal belief in climate change correlated positively ($r = .18, p = .015$) with estimated confidence level, but including this as a covariate in the analysis did not influence the results.

Ratings. Confidence (probability) estimates were positively correlated with ratings of competence ($r = .26, p < .001$), trust ($r = .29, p < .001$), and worry ($r = .16, p = .030$); and negatively correlated with ratings of uncertainty ($r = -.35, p < .001$). In other words, participants who assigned higher confidence levels saw the expert as more competent and trustworthy and less uncertain, and found the statements to cause more worry. Unsurprisingly, personal belief in climate change also correlated positively with ratings of competence ($r = .34, p < .001$), trust ($r = .33, p < .001$), and worry ($r = .40, p < .001$), and correlated negatively with ratings of uncertainty ($r = -.19, p = .009$). There were no significant effects of probability format or interval width on the ratings of the expert (all p 's $> .095$).

Experiment 5

This experiment was run as a replication controlling for some potential confounds in the previous studies. Specifically, a direct comparison of Experiment 1 and 2 was rendered difficult by the fact that these experiments did not only differ in the design, but were also based on different populations (a medium-sized Norwegian student sample in Experiment 1, a large sample of American MTurkers in Experiment 2) and different scenarios. Hence, Experiment 5 was designed as a within-subjects study (like Experiment 1) using a larger sample of MTurkers and the same scenarios as in the other experiments. In addition, participants received a numeracy measure (not included in Experiment 1).

Associative responses, suggesting a link between narrow intervals and high confidence, might be justified if it is assumed that the most confident expert can be more precise because of his or her superior expertise. In the previous experiments, no explicit information was given about the experts' qualifications, allowing participants to draw their own inferences about their expertise.³ As a control, Experiment 5 included a condition where it was explicitly stated that the two experts were of equal expertise.

Method

Participants. American MTurk workers were invited to participate and received a payment of \$1.00 after completing the questionnaire. We excluded 21 participants who spent less than 3 minutes on the survey, failed a simple attention check, or gave erroneous answers (minimum and maximum value the same), leaving 259 participants with a mean age of 36.3 years ($SD = 10.7$), 50.2% were female, and most (80.7%) reported having at least some college education.

³ We thank an anonymous reviewer for raising this point.

Questionnaires. The participants received the same scenarios as in Experiment 2, with the exception that the two different levels of probability/certainty were presented simultaneously (stemming from two different experts) to the same participants in a within-subjects design. Participants were asked to fill in minimum and maximum values for the experts in each of the three scenarios. The order of the high and low probability expert was counterbalanced, and the order of the scenarios was randomized. Probability format (internal vs. external) was varied between subjects. An additional between-subjects factor was the description of the expertise of the forecasters. Participants in one condition were informed that the two experts were both “highly merited researchers who have investigated the consequences of climate change for more than 20 years, and are regarded by their peers as having equal expertise in forecasting environmental changes”. In other words, half of the participants were explicitly informed that the experts were of equal expertise, while there was no mention of this in the other condition (similar to previous experiments). This gives a mixed design with probability level (high vs. low) as within-subjects factor, while probability format (internal vs. external) and expertise (equal vs. not explicitly mentioned) were between-subjects factors.

Participants were not asked to rate the experts in any way in this experiment. Instead, they proceeded to complete the same numeracy test that was used in Experiments 3 and 4.

Results

<Insert Figure 4 about here>

Overall, there was an exactly equal number of responses where 90% intervals were wider than 60% intervals (44.7% of responses) as the other way around (44.7% of responses; the remaining 10.7% responses had 90% and 60% intervals of equal width). As shown in Figure 4, this pattern was not significantly related to probability format (chi-square tests show no difference in either of the scenarios, p 's $> .26$). The responses also did not differ depending on whether participants were explicitly informed that the experts were of equal expertise or not, as confirmed by chi-square tests (p 's $> .20$ for all scenarios). Separate 2 (confidence level: 60% vs. 90%) x 2 (probability format: internal vs. external) x 2 (expertise: equal vs. not explicitly mentioned) mixed ANOVAs for each scenario confirmed this pattern of results: we found no significant main effects (all p 's $> .12$), and the only significant interaction ($p = .050$) was between probability format and expertise for the moth scenario. This interaction may be

coincidental, and did not relate to our hypotheses, so we do not explore it any further. Table 4 shows the descriptives for each scenario.

<Insert Table 4 about here>

Following Experiment 1, we analyzed the subset of respondents who believed that 90% intervals were larger than 60% intervals, and again the median answer in all three scenarios was that 90% intervals were two times larger than 60% intervals. Hence, the “distributional” responses were in line with a normal distribution.

In all three scenarios, the midpoints of the participants’ intervals were higher for the high probability expert (as confirmed by ANOVAs on the raw scores, all p ’s for probability level $< .05$, all other effects, $p > .13$), replicating a finding from the between-subjects version in Experiment 2. This effect seems to be mainly driven by higher maximum values assigned to the 90% expert.

Numeracy. About half of the participants (50.6%) reported to have answered some or all of the numeracy questions before, and scored higher ($M = 4.22$, $SD = 1.76$) than those who indicated they had not seen the test before ($M = 3.37$, $SD = 1.77$), $F(1,257) = 15.131$, $p < .001$, $\eta^2_p = .056$. In this experiment, we found no consistent relationship between numeracy and distributional responses. Overall, the correlation between numeracy and the average standardized interval difference score (the width of the high probability interval minus the width of the low probability interval) was $r = -.04$, $p = .564$.

General Discussion

We started this investigation by asking what people think of the relationship between the width and the probability of a prediction interval. Do they see these two features as compensatory, as required by distributional logic? Or do they rather expect them to be positively associated, from a belief that certainty can manifest itself in both ways, allowing experts to make statements that are both confident and precise? Five studies gave evidence of both kinds of thinking. When participants compare experts that are presented jointly, in a within-subjects design, distributional responses were more frequent than associative responses in Experiment 3 and (to some degree) in Experiment 1. Still, about one third of the participants expected highly confident experts to give more precise predictions (Experiment 1), and conversely, that precise forecasts would be issued with higher confidence (Experiment 3). In Experiment 5, both kinds of responses were equally frequent. Moreover, distributional

responses were not fully compensatory: although the “distributional” respondents suggested that 90% intervals were about twice as large as 60% intervals in Experiments 1 and 5, large differences in interval width entailed much smaller differences in confidence than suggested by a normal probability distribution in Experiment 3. When similar judgments were performed separately, in a between-subjects design, participants expected, on average, the confident expert to be more precise (Experiment 2), and the precise expert to express his predictions with a higher degree of internal (but not external) confidence (Experiment 4). Thus, a joint presentation format may facilitate distributional thinking, whereas ratings performed in a separate mode rather suggests a positive relationship between these two aspects of certainty. Such differences due to presentation format are reminiscent of Hsee’s (1996) analysis of preference reversals between joint and separate evaluations of options (Hsee, Loewenstein, Blount, & Bazerman, 1999).

We do not claim that distributional responses are always correct, as they presuppose that the experts that are compared are (as far as we know) approximately equal in their ability to predict the future. In the present scenarios, they were presented on equal terms, so there was no prior reason to believe that one was superior to the other. However, their predictions might make the more confident one, or the more precise one, *appear* superior. By circular reasoning, one might conclude that this expert will perform better on *both* indicators of expertise, as suggested by the original explanation of the preciseness paradox (Teigen, 1990). Thus the “confidence heuristic” (Price & Stone, 2004), whereby people infer high expertise from high confidence, might be accompanied by a corresponding “preciseness heuristic”, whereby high precision is seen as a cue of high expertise (and hence, of high confidence). However, the association between precision and confidence cannot only be due to perceived expertise, as it persisted in Experiment 5 when the two forecasters were explicitly described as being equally knowledgeable.

Still, we cannot rule out the possibility that high precision and high confidence go together in statements experts actually make. The conversational norms-account of overprecision (Moore et al., 2016) suggests that people try to make their statements as informative as possible. This implies that an expert might try to simultaneously maximize both precision and confidence, at the risk of exaggerating both. Thus we cannot reject “associative” responses as incorrect under all circumstances, even if they seem to defy distributional logic.

The investigation of the role of numeracy gave mixed results. Numeracy appeared unrelated to distributional responses in Experiment 5 where participants generated prediction

intervals corresponding to high or low probabilities. However, when participants were asked to generate probabilities corresponding to narrow or wide intervals (Experiments 3 and 4), numeracy made a difference. In Experiment 3 the difference between wide and narrow intervals was larger for participants with higher numeracy scores. In Experiment 4 those scoring low on numeracy assigned low probabilities to wide intervals, while high numeracy participants perceived wide intervals as much more probable. This suggests that those without sufficient knowledge of statistics intuitively think that vagueness signals lack of knowledge and hence lack of confidence. People who are somewhat more schooled in statistical thinking are able to override this intuitive response. The fact that high and low numeracy participants gave similar ratings to narrow intervals in Experiment 4 also shows that preciseness is generally appreciated. Numerate participants were not more skeptical towards precise forecasts than low numeracy participants (in this context), but more appreciative of the certainty of less precise forecasts. Of course, these findings may not hold under all circumstances. If a statement or forecast is overly precise or overly vague, both low and high numeracy participants may be equally skeptical about its validity.

Probabilities were in all studies expressed in two formats: internal, embedded in a personal statement (I am 90% certain) and external, framed in impersonal language (There is a 90% probability). Experiments 1 and 3 suggested that distributional responses were slightly more common for external statements, but in Experiments 2 and 5, no differences were found. Only Experiment 4 showed a clear effect of probability format, in that external probabilities were higher for wide intervals and internal probabilities were higher for narrow intervals, producing a significant interaction. These results suggest that the framing of a probability statement as internal or external may occasionally influence the way people think about confidence intervals. Specifically, external language may facilitate a more statistical/distributional way of thinking, where imprecise statements give the speaker a higher chance of being (objectively) more correct, whereas internal language is more compatible with associative thinking, where imprecise statements indicate that the speaker is (subjectively) not very confident about his or her own predictions. However, further studies are needed to check whether these results are robust.

Our findings add to the existing literature on confidence intervals, and specifically on overprecision. We did not in the present studies ask people to provide their own intervals, but rather to guess which interval (or which confidence level) another speaker might have in mind. If high confidence levels suggest high precision, it is perhaps not so surprising that people who are asked to provide high confidence intervals (typically 90% or 95%) feel

pressured to be very precise. In any case, they might find it inconsistent to declare themselves highly certain (with a high confidence level) and very uncertain (with wide intervals) at the same time.

The current results have several implications for uncertainty communication. First, it should be useful for experts using confidence intervals to know that recipients with a preference for distributional versus associative thinking may interpret a forecast in diametrically different ways. This is not to say that confidence intervals should be avoided. For instance, confidence intervals in weather forecasts made people more decisive and gave them a better idea of which outcomes to expect compared to point estimates (Savelli & Joslyn, 2013). Information about the full range of potential outcomes gave students better understanding of the life-expectancy of cancer patients than simply the median outcome, which is the usual information given in such contexts (Kirkebøen, 2017). Thus confidence intervals can clearly be useful, but enough information should be provided to ensure that people interpret them in the intended way. An external format, suggesting actual frequencies and variability, will probably convey a more adequate understanding than the upper or lower limits of a person's subjective expectations.

Second, in communicating uncertainty experts should be aware of the compound pressure coming from the expectation of being both confident and precise. This double challenge can be hard to meet without stretching the truth or flat out lying. In a sense, asking for precision and confidence simultaneously is like asking for double overconfidence. An expert who wishes to appear both confident and precise while retaining the uncertainty of her estimate, may choose to provide a single, rather specific boundary instead of both. To say, with high certainty, that "The Eiffel tower is *more than* 250 meters tall" might suggest more expertise than a corresponding, actually more informative statement indicating both minimum and maximum values ("The Eiffel tower is more than 250 and less than 400 meters tall").

The present results are silent about the relative importance of precision and confidence. Is confidence more important than precision, or is it better to be precise than to be confident? Our guess is that people may be less sensitive to interval width than to confidence level, at least in unfamiliar domains. For instance, in forecasts of climate change, it is often not apparent for a layperson what constitutes a wide or a narrow interval. The IPCC states that it is *very likely* that the Atlantic Meridional Overturning Circulation (AMOC) will weaken over the 21st century, with a range for the reduction from 1% to 24% in the best-case scenario (IPCC, 2013). Whether this is a wide or a narrow range is hard to say unless you have more information about the topic. Degree of confidence is more universally evaluable (Hsee &

Zhang, 2010) by having a fixed frame (0-100%). Hence, in unfamiliar domains, we can speculate that it might be more persuasive to have a high confidence level than to indicate a narrow range. The relative importance of confidence and precision is of course an empirical question, which might be sensitive to domain, previous knowledge, personal relevance, and several other factors.

References

- Bagchi, R., & Ince, E. C. (2016). Is a 70% forecast more accurate than a 30% forecast? How level of a forecast affects inferences about forecasts and forecasters. *Journal of Marketing Research*, 53(1), 31-45. doi:10.1509/jmr.12.0526
- Brewer, N., & Burke, A. (2002). Effects of testimonial inconsistencies and eyewitness confidence on mock-juror judgments. *Law and Human Behavior*, 26(3), 353-364. doi:10.1023/A:1015380522722
- Budescu, D. V., & Du, N. (2007). Coherence and consistency of investors' probability judgments. *Management Science*, 53(11), 1731-1744. doi:10.1287/mnsc.1070.0727
- Cokely, E. T., Galesic, M., Schulz, E., Ghazal, S., & Garcia-Retamero, R. (2012). Measuring risk literacy: The Berlin Numeracy Test. *Judgment and Decision Making*, 7(1), 25-47.
- Connolly, T., & Dean, D. (1997). Decomposed versus holistic estimates of effort required for software writing tasks. *Management Science*, 43(7), 1029-1045. doi:10.1287/Mnsc.43.7.1029
- Du, N., Budescu, D. V., Shelly, M. K., & Omer, T. C. (2011). The appeal of vague financial forecasts. *Organizational Behavior and Human Decision Processes*, 114(2), 179-189. doi:10.1016/j.obhdp.2010.10.005
- Fox, C. R., & Ülkümen, G. (2017). Comment on Løhre & Teigen. *Thinking & Reasoning*.
- Gillies, D. (2000). *Philosophical theories of probability*. London: Routledge.
- Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics, Vol. 3: Speech acts* (pp. 41-58). New York: Academic Press.
- Hacking, I. (1975). *The emergence of probability: A philosophical study of early ideas about probability, induction and statistical inference*. Cambridge, UK: Cambridge University Press.
- Hayward, M., & Fitza, M. (2016). Pseudo-precision? Precise forecasts and impression management in managerial earnings forecasts. *Academy of Management Journal*. doi:10.5465/amj.2014.0304
- Hsee, C. K. (1996). The evaluability hypothesis: An explanation for preference reversals between joint and separate evaluations of alternatives. *Organizational Behavior and Human Decision Processes*, 67(3), 247-257. doi:10.1006/Obhd.1996.0077
- Hsee, C. K., Loewenstein, G. F., Blount, S., & Bazerman, M. H. (1999). Preference reversals between joint and separate evaluations of options: A review and theoretical analysis. *Psychological Bulletin*, 125(5), 576-590. doi:10.1037//0033-2909.125.5.576

- Hsee, C. K., & Zhang, J. A. (2010). General evaluability theory. *Perspectives on Psychological Science*, 5(4), 343-355. doi:10.1177/1745691610374586
- IPCC (2013). Summary for policymakers. In T. F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, & P. M. Midgley (Eds.), *Climate change 2013: The physical science basis. Contribution of Working Group I to the fifth assessment report of the Intergovernmental Panel on Climate Change*. Cambridge, UK, and New York, USA: Cambridge University Press.
- Janiszewski, C., & Uy, D. (2008). Precision of the anchor influences the amount of adjustment. *Psychological Science*, 19(2), 121-127. doi:10.1111/J.1467-9280.2008.02057.X
- Jerez-Fernandez, A., Angulo, A. N., & Oppenheimer, D. M. (2014). Show me the numbers: Precision as a cue to others' confidence. *Psychological Science*, 25(2), 633-635. doi:10.1177/0956797613504301
- Jørgensen, M. (2016). The use of precision of software development effort estimates to communicate uncertainty. In D. Winkler, S. Biffl, & J. Bergsmann (Eds.), *Software Quality. The Future of Systems- and Software Development* (pp. 156-168). Cham: Springer International Publishing.
- Kahneman, D., & Tversky, A. (1982). Variants of uncertainty. *Cognition*, 11(2), 143-157. doi:10.1016/0010-0277(82)90023-3
- Keren, G., & Teigen, K. H. (2001a). The probability-outcome correspondence principle: A dispositional view of the interpretation of probability statements. *Memory & Cognition*, 29(7), 1010-1021. doi:10.3758/Bf03195763
- Keren, G., & Teigen, K. H. (2001b). Why is $p = .90$ better than $p = .70$? Preference for definitive predictions by lay consumers of probability judgments. *Psychonomic Bulletin & Review*, 8(2), 191-202. doi:10.3758/Bf03196156
- Kirkeboen, G. (2017). "The median isn't the message" (1985) - thirty years later: How to communicate prognostic uncertainty to patients. Manuscript submitted for publication.
- Langnickel, F., & Zeisberger, S. (2016). Do we measure overconfidence? A closer look at the interval production task. *Journal of Economic Behavior & Organization*, 128, 121-133. doi:10.1016/j.jebo.2016.04.019
- Loschelder, D. D., Stuppi, J., & Trötschel, R. (2014). "€14,875?!": Precision boosts the anchoring potency of first offers. *Social Psychological and Personality Science*, 5(4), 491-499. doi:10.1177/1948550613499942

- Løhre, E. (2017). *Stronger, sooner, and more certain climate change: A link between certainty and outcome strength in revised forecasts*. Manuscript submitted for publication.
- Løhre, E., & Jørgensen, M. (2016). Numerical anchors and their strong effects on software development effort estimates. *Journal of Systems and Software*, 116, 49-56. doi:10.1016/j.jss.2015.03.015
- Løhre, E., & Teigen, K. H. (2016). There is a 60% probability, but I am 70% certain: Communicative consequences of external and internal expressions of uncertainty. *Thinking & Reasoning*, 22(4), 369-396. doi:10.1080/13546783.2015.1069758
- Mason, M. F., Lee, A. J., Wiley, E. A., & Ames, D. R. (2013). Precise offers are potent anchors: Conciliatory counteroffers and attributions of knowledge in negotiations. *Journal of Experimental Social Psychology*, 49(4), 759-763. doi:10.1016/J.Jesp.2013.02.012
- McKenzie, C. R. M., & Amin, M. B. (2002). When wrong predictions provide more support than right ones. *Psychonomic Bulletin & Review*, 9(4), 821-828. doi:10.3758/Bf03196341
- McKenzie, C. R. M., Liersch, M. J., & Yaniv, I. (2008). Overconfidence in interval estimates: What does expertise buy you? *Organizational Behavior and Human Decision Processes*, 107(2), 179-191. doi:10.1016/J.Obhdp.2008.02.007
- Miller, D. J., Spengler, E. S., & Spengler, P. M. (2015). A meta-analysis of confidence and judgment accuracy in clinical decision making. *Journal of Counseling Psychology*, 62(4), 553-567. doi:10.1037/cou0000105
- Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*, 115(2), 502-517. doi:10.1037/0033-295x.115.2.502
- Moore, D. A., Tenney, E. R., & Haran, U. (2016). Overprecision in judgment. In G. Wu & G. Keren (Eds.), *Handbook of Judgment and Decision Making*. New York: Wiley.
- Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science*, 23(3), 184-188. doi:10.1177/0963721414531598
- Price, P. C., & Stone, E. R. (2004). Intuitive evaluation of likelihood judgment producers: Evidence for a confidence heuristic. *Journal of Behavioral Decision Making*, 17(1), 39-57. doi:10.1002/bdm.460
- Read, C. (1920). *Logic: Deductive and inductive* (4th ed.). London: Simpkin, Marshall, Hamilton, Kent & Co., Ltd.

- Sah, S., Moore, D. A., & MacCoun, R. J. (2013). Cheap talk and credibility: The consequences of confidence and accuracy on advisor credibility and persuasiveness. *Organizational Behavior and Human Decision Processes*, 121(2), 246-255. doi:10.1016/J.Obhdp.2013.02.001
- Savelli, S., & Joslyn, S. (2013). The advantages of predictive interval forecasts for non-expert users and the impact of visualizations. *Applied Cognitive Psychology*, 27(4), 527-541. doi:10.1002/acp.2932
- Schwartz, L. M., Woloshin, S., Black, W. C., & Welch, H. G. (1997). The role of numeracy in understanding the benefit of screening mammography. *Annals of Internal Medicine*, 127(11), 966-972.
- Snizek, J. A., & Van Swol, L. M. (2001). Trust, confidence, and expertise in a Judge-Advisor System. *Organizational Behavior and Human Decision Processes*, 84(2), 288-307. doi:10.1006/obhd.2000.2926
- Teigen, K. H. (1990). To be convincing or to be right: A question of preciseness. In K. J. Gilhooly, M. T. G. Keane, R. H. Logie, & G. Erdos (Eds.), *Lines of Thinking* (Vol. 1). Chichester: Wiley.
- Teigen, K. H., & Jørgensen, M. (2005). When 90% confidence intervals are 50% certain: On the credibility of credible intervals. *Applied Cognitive Psychology*, 19(4), 455-475. doi:10.1002/Acp.1085
- Teigen, K. H., & Løhre, E. (2017). Expressing (un)certainly in no uncertain terms: Reply to Fox and Ülkümen. *Thinking & Reasoning*. doi:10.1080/13546783.2017.1314965
- Tenney, E. R., MacCoun, R. J., Spellman, B. A., & Hastie, R. (2007). Calibration trumps confidence as a basis for witness credibility. *Psychological Science*, 18(1), 46-50. doi: 10.1111/j.1467-9280.2007.01847.x
- Yaniv, I., & Foster, D. P. (1995). Graininess of judgment under uncertainty: An accuracy-informativeness trade-off. *Journal of Experimental Psychology: General*, 124(4), 424-432. doi:10.1037//0096-3445.124.4.424
- Yaniv, I., & Foster, D. P. (1997). Precision and accuracy of judgmental estimation. *Journal of Behavioral Decision Making*, 10(1), 21-32. doi:10.1002/(Sici)1099-0771(199703)10:1<21::Aid-Bdm243>3.0.Co;2-G
- Yates, J. F., Price, P. C., Lee, J. W., & Ramirez, J. (1996). Good probabilistic forecasters: The 'consumers' perspective. *International Journal of Forecasting*, 12(1), 41-56. doi:10.1016/0169-2070(95)00636-2

- Ülkümen, G., Fox, C. R., & Malle, B. F. (2016). Two dimensions of subjective uncertainty: Clues from natural language. *Journal of Experimental Psychology: General*.
- Zhang, Y. C., & Schwarz, N. (2013). The power of precise numbers: A conversational logic analysis. *Journal of Experimental Social Psychology*, 49(5), 944-946.
doi:10.1016/J.Jesp.2013.04.002

Tables

Table 1. *Mean estimated interval widths of low and high probability statements in three scenarios, between-subjects, Experiment 2 (standard deviations in parentheses).*

Scenario	Internal probability format		External probability format	
	60% (<i>n</i> = 59)	90% (<i>n</i> = 62)	60% (<i>n</i> = 57)	90% (<i>n</i> = 64)
1. Trout population	20.6 (11.1)	19.5 (15.5)	22.9 (13.2)	17.3 (10.7)
2. Moth area	26.9 (17.6)	22.6 (16.4)	26.5 (17.4)	21.9 (12.4)
3. Silver fir expansion	66.2 (43.5)	62.5 (43.3)	64.5 (38.6)	54.4 (39.1)
Average (log- and z-transformed)	.11 (.80)	-.06 (.84)	.17 (.77)	-.20 (.97)

Table 2. *Mean estimated confidence levels of two experts differing in interval width in three scenarios, within-subjects, Experiment 3 (standard deviations in parentheses).*

Scenario	Internal probability format (<i>n</i> = 100)		External probability format (<i>n</i> = 102)	
	Narrow	Wide	Narrow	Wide
1. Trout population	70.8 (22.3)	72.9 (23.5)	56.0 (25.2)	61.4 (27.6)
2. Moth area	74.1 (17.9)	75.9 (18.4)	63.9 (19.0)	69.4 (20.2)
3. Silver fir expansion	75.0 (17.4)	76.3 (20.7)	66.9 (21.9)	72.6 (21.2)
Average	73.3 (17.1)	75.1 (18.9)	62.3 (18.8)	67.8 (19.7)

Table 3. *Mean estimated confidence levels of wide and narrow interval statements in three scenarios, between-subjects, Experiment 4 (standard deviations in parentheses).*

Scenario	Internal probability format		External probability format	
	Narrow (<i>n</i> = 49)	Wide (<i>n</i> = 48)	Narrow (<i>n</i> = 49)	Wide (<i>n</i> = 46)
1. Trout population	75.6 (18.6)	65.1 (23.9)	60.8 (26.1)	67.9 (25.6)
2. Moth area	77.0 (14.9)	71.2 (18.4)	68.4 (20.9)	74.4 (17.9)
3. Silver fir expansion	75.0 (18.2)	71.8 (19.2)	68.3 (23.9)	73.0 (20.4)
Average	75.9 (15.0)	69.3 (18.3)	65.9 (21.4)	71.8 (18.7)

Table 4. *Mean estimated interval widths of two experts differing in confidence level in three scenarios, within-subjects, Experiment 5 (standard deviations in parentheses). The difference score was computed by subtracting the width of the 60% expert from that of the 90% expert.*

Scenario	Internal probability format (<i>n</i> = 133)		External probability format (<i>n</i> = 126)	
	60%	90%	60%	90%
1. Trout population	17.9 (11.8)	20.0 (15.8)	17.8 (11.7)	19.0 (16.7)
2. Moth area	21.5 (13.5)	25.1 (20.7)	22.0 (14.0)	22.2 (17.1)
3. Silver fir expansion	47.5 (36.1)	52.1 (41.7)	47.3 (38.9)	46.3 (42.4)
Average difference score, z-transformed	.06 (.84)		-.06 (.93)	

Figure captions

Figure 1. *Percentages of participants making high probability (90%) intervals wider, equal to, or narrower than low probability (60%) intervals, Experiment 1.*

Figure 2. *Percentages of participants assigning higher, lower, or equal probabilities to wide and narrow interval statements, Experiment 3.*

Figure 3. *Mean estimated probability levels for wide and narrow interval statements by low and high numeracy participants, Experiment 4.*

Figure 4. *Percentages of participants making high probability (90%) intervals wider, equal to, or narrower than low probability (60%) intervals, Experiment 5.*