Organizational Behavior and Human Decision Processes 171 (2022) 104166

Contents lists available at ScienceDirect

Organizational Behavior and Human Decision Processes

journal homepage: www.elsevier.com/locate/obhdp

What is a "likely" amount? Representative (modal) values are considered likely even when their probabilities are low

Karl Halvor Teigen [a,*], Marie Juanchich [b], Erik Løhre [c]

[a] University of Oslo, Norway
[b] University of Essex, UK
[c] BI Norwegian Business School, Oslo, Norway

ARTICLE INFO

Keywords:
Likely
Verbal probabilities
Subjective probability
Graph literacy
Representativeness
Uncertainty communication
Cost estimates

ABSTRACT

Research on verbal probabilities and standard scales issued by national and international authorities suggest that only events with probabilities above 60% should be labelled "likely". We find, however, that when people apply this term to continuous variables, like expected costs, it describes the *most* likely (modal) outcome or interval, regardless of actual probabilities, which may be quite small. This was demonstrated in six studies in which lay participants ($N = 2{,}228$) were shown probability distributions from various domains and asked to generate or to select "likely" outcome intervals. Despite having numeric and graphically displayed information available, participants judged central, low-probability segments as "likely" (as opposed to equal or larger segments in the tails) and subsequently overestimated the chances of these outcomes. We conclude that high-probability interpretations of "likely" are only valid for binary outcomes but not for distributions of graded variables or multiple outcomes.

1. Introduction

Uncertainty comes in many flavors and can be found everywhere in our private and professional lives. People can be uncertain about a medical diagnosis, the costs of a highway project, the results of the next (and indeed, of the past) presidential election, and the outcome of tomorrow's football match. Following a recent taxonomy (van der Bles et al., 2019), we can distinguish between *what* we are uncertain about (the object), and *how* to express it (the format). The objects of uncertainty can be *facts* (categorical outcomes), *numbers* (continuous variables), and *hypotheses* (propositions). Uncertainties vary in degrees, which can be expressed and communicated in a *verbal* format, by terms like "unlikely", "possible" and "likely", in a *numeric* format, as probabilities, percentages, or intervals, and *visually* by graphs depicting a probability distribution. Much research has been devoted to comparisons between the verbal and numeric formats, which one is more preferred (Erev & Cohen, 1990; Juanchich & Sirota, 2020), more understandable (Wallsten et al., 1993), more normative (Windschitl & Wells, 1996), or more efficient (Budescu et al.,1988; Collins & Mandel, 2019; Mandel & Irwin, 2021), and indeed how they should be coordinated and translated into each other (e.g., Budescu & Wallsten, 1995; Budescu et al., 2014; Clark, 1990; Dhami & Mandel, 2022; Wallsten et al., 1993; Wintle et al., 2019).

The present research extends past work by exploring the way we talk about an understudied object of uncertainty. While past research has studied format effects in a context of binary categorical outcomes, we investigate the meaning of a basic verbal term, "likely", when used to describe the outcomes of a continuous variable, i.e., numbers, quantities, amounts and magnitudes of objects that can be counted or measured, as opposed to what it supposedly means when used to describe uncertainties about a dichotomous fact. In many contexts dichotomization is not appropriate, nor informative, because decision-makers need a finer grained level of information, focusing on which quantities that are expected. What are the likely costs of a proposed project? What is a likely rise of ocean level in the future? What is a likely number of people that will be contaminated with a new variant of the Corona virus? All these questions are about continuous variables rather than dichotomies, which can, in principle be portrayed as a probability density distribution, or a multi-categorical probability function like those presented in Fig. 1.

Previous studies of verbal probability expressions (VPE) have typically concluded that verbal terms are vague and imprecise compared to numbers, but that they can be associated with identifiable segments of the 0–1 probability scale (Budescu & Wallsten, 1995; Collins & Hahn,

* Corresponding author at: Department of Psychology, University of Oslo, Norway.
E-mail address: k.h.teigen@psykologi.uio.no (K.H. Teigen).

https://doi.org/10.1016/j.obhdp.2022.104166
Received 8 June 2021; Received in revised form 21 April 2022; Accepted 17 May 2022
Available online 7 June 2022

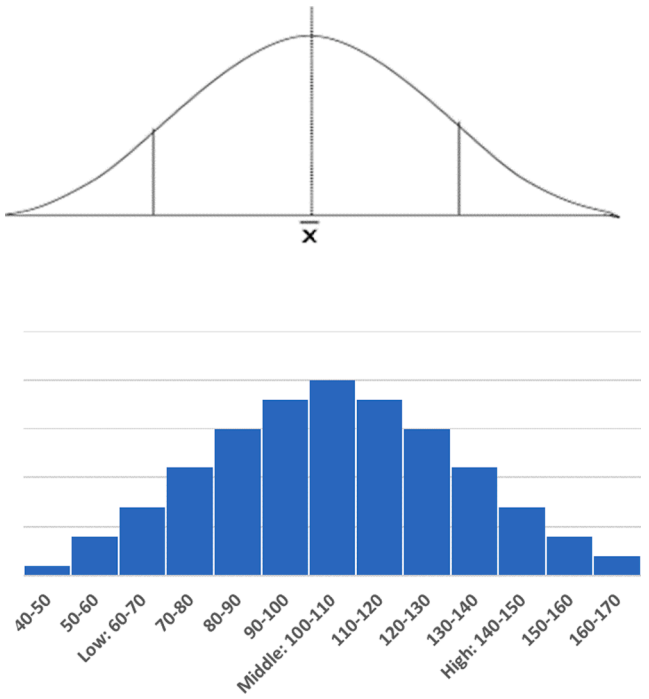


Fig. 1. Graphs illustrating probability distributions of expected costs. Upper panel: Smooth curve used in Study 1 (adapted from Teigen et al., 2020). Lower panel: Stepwise (binned) distribution used in Study 2, 4, and 5.

2018; Juanchich et al., 2019). Outcomes or estimates that are "virtually certain" have a probability close to 1 (or 100%), whereas those that are "unlikely" are only 20–30% probable. "Likely" (and its close synonym: "probable") is deemed to correspond, on average, to probabilities around 70%, and in any case higher than 50% (e.g., Budescu & Wallsten, 1995; Clark, 1990; Lichtenstein & Newman, 1967; O'Brien, 1989; Sirota & Juanchich, 2015; Theil, 2002).

"Likely" also holds a central place in recommended scales for conveying uncertainty in professional disciplines, including climate science, medicine, military intelligence, food safety, and risk analysis, often with modifiers ("very likely", "extremely likely", "more likely than not"). Such scales propose that "likely" (without a modifier) should be used for describing outcomes that have an estimated probability of 66–90% (European Food Safety Authority, 2019; The International Panel of Climate Change, Mastrandrea et al., 2010), 55–80% (US Intelligence, ICD, 2015), or 60–90% (Irwin & Mandel, 2020; North Atlantic Treaty Organization, 2016).

However, these translations and guidelines appear to have as their object uncertain facts rather than uncertain numbers. When speakers claim that "likely" means $p = 70\%$, they typically have a binary rather than a continuous outcome in mind, and are evaluating whether a particular effect, or class of effects (like health damages from smoking, or more frequent floods from climate change) can be expected to occur or not occur. This binary interpretation can be extended to cases where the underlying dimension is of a quantitative and continuous nature, but can be dichotomized (for instance, in discussions of whether the costs of a project will stay within the budget limits or not, or whether global temperatures will exceed or not exceed a target value).

The *binary assumption* holds for translations of other standard terms, although this is not explicitly stated. For instance, when "even chance" is supposed to mean $p = 50\%$, as in the NATO and ICD guidelines, it obviously presupposes a binary event. If several (i.e., more than two) players have an "even chance" of winning, they cannot have a 50% chance each. It follows that the standard translations of VPEs might be misleading for predictions of multiple outcomes.

Some prior studies indicate that people sometimes describe low-probability outcomes in a set of multiple discrete outcomes as being probable when they appear to have the same chances of occurring, even if the mean chances of n outcomes cannot exceed $1/n$ ("the equiprobability effect", Teigen, 1988; 2001). This effect could be due to a causal "propensity" model of probabilities. A runner in the lead group is in position to get in front of others and could "easily" gain the extra inch needed for winning. A lottery player, with no control of the outcome, might still have "a good chance" of winning simply because there is no obstacle to prevent this specific number from being drawn (Teigen, 2001). An outcome may also be called "likely" when it has a better chance than its closest competitor ("the alternative outcomes effect", Windschitl & Wells, 1998). In these studies, the term "likely" appears to be used in a comparative rather than in an absolute sense.

1.1. *Two paradigms for "likely"*

The main focus of the present research is on continuous (more accurately: graded) variables rather than discrete outcomes, like winning vs. not winning a prize or a competition. To explore the pragmatic meaning of selected VPEs in a context of quantities it may be helpful to rephrase the question. In addition to asking which probabilities correspond to selected VPEs, one might ask *which outcomes* in a distribution should be described with these VPEs. This alternative way of identifying the meaning of VPEs was called the "which outcome-approach" (Teigen et al., 2014), and has led to several surprising findings, deviating from the meanings established by the conventional "translation" approach. For instance, while "possible" is commonly supposed to reflect a middle probability (around $p = .5$), it is used in a context of quantities to denote extreme, top outcomes ("a temperature rise of 5 degrees is possible"; Teigen et al., 2018). But such outcomes have typically a very low probability of occurring (Juanchich et al., 2013; Teigen et al., 2020). "Unlikely", which is assumed to correspond to numeric probabilities around 20%, will prompt participants to select an outcome beyond the maximum of an outcome distribution, with probabilities close to zero (Jenkins et al., 2018; Teigen et al., 2013). For instance, participants who were shown a distribution of the duration of batteries, ranging from 1.5 to 3.5 h, found it most appropriate to say that a (maximal) duration of 3.5 h is *possible* and 4 h (out of range) is *unlikely*.

In studies based on this approach, people were willing to call some outcomes "likely" or "probable" that had a probability of less than 50% to occur (Teigen et al., 2013; Teigen et al., 2014). For instance, a C grade was claimed to be "probable" at an exam where only 40 out of 100 candidates obtained that grade (Teigen et al., 2013, Experiment 4). However, these distributions contained no alternative that had a probability of more than 50%, which may have induced participants to choose the most likely outcome available, and/or to consider the selected alternative as representing a broader category (e.g., medium grades) which jointly had a high chance of occurring. These interpretations were controlled for in the present studies by asking for a freely chosen *range* instead of point values. A range can be widened to contain the larger part of a distribution if so desired.

The present studies were designed to contrast two potential interpretations of a "likely" event: one formal, based on proportions of the outcome space, the other pragmatic, based on word usage.

According to a *formal,* probabilistic approach, reflected in traditional translation studies and standard prescriptive scales, "likely" signifies high probabilities from 60% and upwards. This mathematical or technical definition would lead to outcome ranges which should be wide enough to incorporate the greater part of a distribution.

According to a more colloquial, *pragmatic* definition, a "likely" outcome might simply mean one we should be prepared for and expect more strongly than other, more deviant outcomes. A "likely" range would be one containing the most typical or frequent instances of a distribution. This range should exclude divergent outcomes that are not, in Kahneman and Tversky's (1972) terminology, "representative", and could accordingly be rather narrow. This interpretation rests on

conversational principles or maxims of communication (Grice, 1975), according to which a message should contain an optimal amount of information. An estimate of the "likely" costs of a construction project to fall between 0 and 200 million is technically highly probable, but such an estimate would be too wide to be informative (Du et al., 2011; Løhre & Teigen, 2017; Yaniv & Foster, 1995).

These two interpretations of likely could lead to similar results when applied to binary partitions. In the distribution of costs displayed in the lower part of Fig. 1, costs of "more than 80 million" are *likely,* both in terms of probabilities (this section includes the greater part of the curve) and in terms of expectations (this section includes the average expected costs), whereas the complementary set of "less than 80 million" is *not* likely, neither in a technical, nor in a pragmatic sense (it contains less than half of the curve, and includes mostly deviant, unexpected values). However, for the full spectrum of binned outcomes from low to high the technical and pragmatic definitions differ. The peak outcome contains much less than 60% of the distribution and should not formally be defined as "likely", but on the other hand it is a representative outcome and thus "likely" from a pragmatic perspective.

1.2. The present studies

In the present work, we evaluate whether people use "likely" in the formal probabilistic sense (likely = a high probability) or rather in a more pragmatic way (likely = a "representative" value). We probe the probabilistic and pragmatic usage of this concept using two complementary designs, using likely as an independent variable or a dependent variable, respectively, asking participants either to design a likely interval, or giving them an interval and asking them whether it is likely or not.

In Study 1–3 we ask participants about the width of *likely* compared to *most likely* intervals. From a formal point of view, the *most likely* interval could be very narrow, perhaps limited to one single, middle value, whereas a *likely* interval should be wide enough to encompass 60–80% of the distribution. We further ask participants to estimate the numeric probabilities of both these intervals. This should, again formally, reflect the corresponding areas under the curve (60–80% for sufficiently wide intervals and perhaps down to 15–20% for very narrow ones). From a pragmatic perspective, guided by representativeness, a likely and a most likely interval can be more similar, and the numeric probabilities attached to these segments may differ from their proportions in the distribution.

In Study 4–6 we present segments of a probability distribution and ask whether these segments can be considered likely or not. From a formal point of view, this should depend on their size (how much of the distribution they cover), rather than their placement as central or peripheral, so that a wide segment covering more than half of the curve might be called likely, whereas one that cover less than 50% must be called something else (e.g., "not likely"). From a pragmatic view a segment may be considered likely if centrally located in the distribution or including peak values.

The judgment tasks were in all studies supported by graphs, showing the distribution of a continuous outcome that we manipulated across studies. Study 1 displayed a smooth, continuous curve without numbers along the axes, with exception of the middle value. In Study 2 the estimation task was made more transparent by binning outcomes into bars with specified values along the x-axis. In Study 3 the areas were still more precisely defined by labelling all bars with corresponding percentages. In this study we also added a within-subjects condition, allowing the participants to make an explicit choice of whether *likely* and *most likely* estimates should be considered equal, or be given different interpretations. We further asked participants in these studies to "back-translate" the intervals they had selected into numeric probabilities. In Studies 4–6 participants received selected partitions of a probability distribution (wide or narrow, central or peripheral) and were asked which of them could be considered "likely". Study 6 differed from the other studies by presenting a skewed distribution, where the modal (peak) outcome differed from the median, to explore how shape of distribution would affect the identification of a "likely" outcome. This study also compared outcomes along a graded (continuous) dimension with a set of multiple categorical (discrete) outcomes. As probability judgments have in the past been related to numeracy and amount of formal schooling (Lipkus et al., 2001; Lipkus & Peters, 2009; Reyna et al., 2009), Studies 1, 2, 3, and 5 included level of education as a background variable, and Study 3a and 3b included a graph literacy scale.

Preregistrations of Studies 2–6 and data files for the results of all the studies are available at https://osf.io/ueqs9/?view_only=089d6b1b76bc40ff81a22fa627323096.

2. Study 1

In Studies 1–3 we asked participants about what is a *likely* and what is the *most likely* outcome, based on a probability density function of costs for a large construction project. The vignette was derived from an actual case in the quality assurance documents for a Norwegian road construction project (adapted from Teigen et al., 2020). Current principles of governance of large public investments require calculations of a complete cumulative probability distribution of costs, where estimates corresponding to P15, P50 and P85 are given special attention and often described as minimum, most likely, and maximum estimates (Volden & Andersen, 2018; Volden & Samset, 2017). P50 (the median) is in these reports variously described as "an expected" or "most probable" value, and sometime even called a "likely" estimate, despite the fact that this exact value obviously has a low probability of occurrence. This terminological mix of phrases could give rise to misunderstandings about the estimates' probabilistic meanings and give recipients the impression that "likely" and "most likely" mean roughly the same.

2.1. Method

Participants. Participants were recruited online from UK and Ireland via the crowdsourcing platform Prolific. They received the questions about likelihood after a brief, unrelated questionnaire. After excluding participants who failed a simple attention check or spent less than one minute on the whole survey, the final sample consisted of 220 participants (156 women, 63 men, and 1 other), with ages ranging from 18 to 80 years (M = 33.8, SD = 12.0). Almost half of them (48.6%) reported to have completed higher education, corresponding to a bachelor's degree or higher.

Material and procedure. All questionnaires contained a brief description about the quality assurance procedure of large public projects in Norway. Participants in the experimental conditions were told that an independent expert team had calculated a middle estimate of expected costs for the recommended alternative along with a low and a high estimate. The meaning of these values was briefly described and illustrated with a symmetric bell-shaped curve for a real highway project where P15, P50, and P85 were clearly marked, as shown by the graph in the upper panel of Fig. 1. The middle estimate was stated to be £110 million (originally NOK 1100 million), whereas the numeric estimates corresponding to P15 and P85 were not disclosed (they were in the original model estimated to be NOK 750 million and NOK 1450 million, respectively).

Participants in two Experimental conditions were asked to imagine that the expert team was asked about the *likely* [*most likely*] costs of this road project, by completing this statement:

Likely [*Most likely*] costs will be between …… million and ….. million.

They were then told that the experts were asked to indicate the probability of these "likely" ["most likely"] costs on a scale from 0 to 100%.

Participants in two Control conditions were not shown the graph, but

Table 1
Mean intervals (in million pounds) and probability estimates (0–100%) for predicted costs that were selected by or assigned to participants as being "likely" or "most likely", Studies 1–3. (Standard deviations in parentheses.)

Study Condition	Interval width		Subjective probabilities	
	Likely	Most likely	Likely	Most likely
Study 1				
Experimental (subjective intervals)	44.8 (44.6)	45.2 (38.9)	61.7% (24.1)	66.1% (20.4)
Control (assigned intervals)	30.0 (0.0)	30.0 (0.0)	72.8% (15.6)	75.1% (17.4)
Study 2				
Experimental (subjective intervals)	£45.0 (33.5)	£37.6 (29.4)	69.2% (17.0)	70.8% (16.9)
Control (assigned intervals)	£30.0 (0.0)	£30.0 (0.0)	71.0% (16.6)	74.8% (16.9)
Study 3				
3A: Between subjects	£35.5 (25.6)	£35.8 (30.1)	54.1% (26.4)	48.2% (29.6)
3B: Within subjects	£35.3 (24.8)	£24.1 (22.1)	56.9% (26.4)	52.5% (29.7)

simply told that the expert team had estimated costs between £100 million and £120 million to be *likely* or *most likely*. This assigned interval was deliberately chosen to be smaller than the original 70% range. They were subsequently asked to estimate the probability equivalents of these expressions, as above. For a full description of materials, see Appendix A.

2.2. Results

Participants in the two experimental conditions suggested on average that the expert team would describe costs between 86.6 million ($SD = 21.8$) and 132.4 million ($SD = 31.1$) as "likely", and costs between 91.1 million ($SD = 18.5$) and 136.3 million ($SD = 31.4$) as "most likely". There were no significant differences between these two sets of estimates. We found accordingly no evidence for "likely" to be used about a wider range of outcomes than "most likely", in fact the ranges in the two experimental conditions were almost identical, as shown in the upper panel of Table 1.

The numeric probabilities suggested for these two ranges were also highly similar (see Table 1). Interestingly, the intervals produced by participants in the experimental conditions were considered less probable than the much narrower intervals assigned to them in the control groups. A 2 × 2 ANOVA with Group (Experimental vs. Control) and VPE (Likely vs. Most likely) as the two independent factors revealed a significant difference for group, $F(1, 216) = 14.57$, $p < .001$, but none for VPE, $F(1, 216) = 1.57$, $p = .211$, and no significant interaction $F(1, 216) = 0.180$, ns.

Interval and probability judgments were not related to level of education. Separate ANOVAs for interval width and probability estimates revealed no differences between participants with higher (bachelor's degree or more) vs. lower level of academic education (see Tables D1 and D2 in Appendix D).

2.3. Discussion

Participants in the present study underestimated the uncertainty ranges in the original report (as cited in Teigen et al., 2020), where the middle 70% had been calculated as an interval spanning a 70 million rather than a 45 million interval. This is in line with the "over-precision" of range judgments reported in previous research (Moore & Healy, 2008; Peterson & Pitz, 1988; Teigen et al., 2020). More important: they did not distinguish between "likely" and "most likely" estimates, neither in terms of ranges nor in terms of probabilities. With smaller assigned intervals in the control groups, participants felt that the predictions were more (rather than less) likely. With self-provided ranges, the correlations between interval estimates and subjective probability estimates were close to zero ($r = -0.08$ and $r = 0.06$ in the Likely and Most likely conditions, respectively). Formally, one should expect a narrow interval to capture a smaller proportion of the distribution, but previous research (Løhre & Teigen, 2017; Løhre et al., 2019) has shown that lay people are divided on this issue.

A failure to distinguish between "likely" and "most likely" costs could lead to grave misunderstandings. For example, readers of quality assurance reports might incorrectly assume that a median cost estimate of 110 milllion is more likely than not ($p > .5$). This misinterpretation would be further propagated by communicators who use "likely" and "most likely" interchangeably about the same value. The middle estimate is occasionally referred to as the "expected" value, with no distinction being made between "expected" and "most expected". As a matter of fact, it would be quite unusual that the costs should turn out exactly as "expected".

The continuous curve used in Study 1 with no units on the x-axis made it difficult to assess the magnitude of different areas under the curve, and may have concealed their potential relevance for probability assessments. It also obscured an evaluation of the accuracy of assessments. Participants who felt that "most likely" corresponded to a point may have been puzzled by being required to generate an interval. In the following studies the distributions were partitioned (binned) into 10 million segments, with proportions reflected by the height of bars. This erases the distinction between point and intervals, as the "most likely" point estimate is equal to the 10-million interval of the tallest (middle) bar.

3. Study 2

This study was a preregistered replication of Study 1 (AsPredicted reference #58641), with two important changes: (1) The smooth curve illustrating the probability distribution was redesigned into to a bar graph, with the height of bars representing the probabilities of 13 adjacent intervals along the curve. (2) The x-axis was labelled from 40 million to 170 million in 10 million increments. These two features would enable more precise estimates both of the ranges and of the associated probabilities, although exact information about the height of the bars was not provided and had to be judged from visual inspection of the graph.

3.1. Method

Participants. Participants were recruited from the UK and Ireland using Prolific. They received the estimation tasks appended to an unrelated questionnaire. After excluding five participants who gave incomplete or ambiguous answers, the final sample consisted of 465 participants (323 women, 135 men, 7 other), with ages ranging from 18 to 74 years ($M = 34.5$, $SD = 11.9$). Almost two thirds (65.3%) reported having obtained a bachelor's degree or equivalent. They were randomly assigned to one of four conditions, two experimental and two control groups.

Material and procedure. All questionnaires contained a brief description about the quality assurance procedure involved in the planning of large public projects, with only minor changes from the description used in Study 1. Participants in the experimental conditions were told that an expert team had calculated the expected costs as a

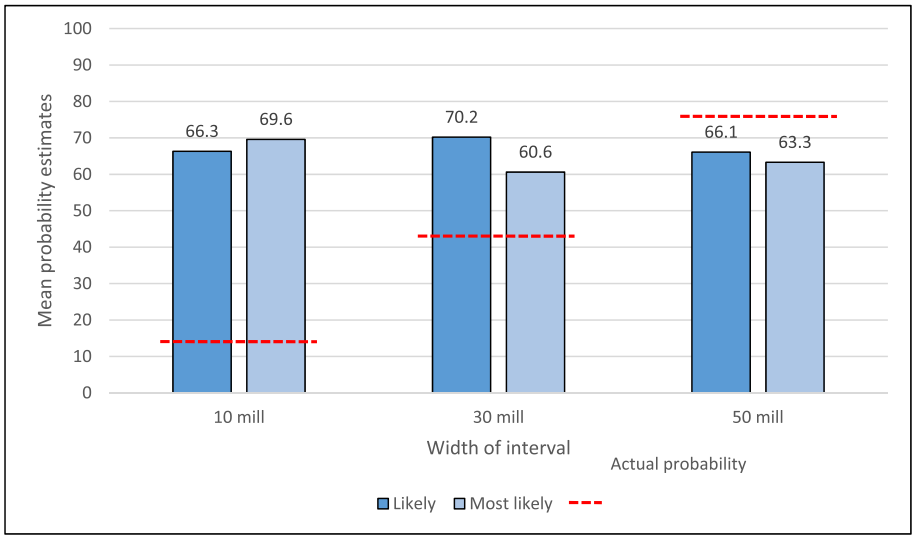


Fig. 2. Mean subjective probabilities of "likely" and "most likely" ranges of different widths: 10 (100–110) million, 30 (90–120) million, and 50 (80–130) million. Dashed lines: Correct probabilities based on graph.

probability distribution, around a middle estimate (in this case £105 million), along with lower and higher estimates that were less likely to occur.

The entire probability distribution was presented as a bar graph, as shown in the lower part of Fig. 1.

The use of bars was introduced to facilitate estimates of probabilities as areas under the curve. For instance, the probability of middle values in the 100–110 interval is approximately 15%, the 90–120 interval (the three middle bars) contains about 42% of the curve, and so on.

Participants in two Experimental conditions were asked to imagine that the expert team was asked about the *likely* [*most likely*] costs, by filling in the missing figures in this statement:

Likely [Most likely] costs will be between …… million and ….. million.

They were subsequently informed that the experts had stated the probability of these "likely" ["most likely"] costs on a scale from 0 to 100 percent, in this statement: "By the expression "likely" ["most likely"] costs we mean costs that have a probability of about ….% of occurring".

Participants in two Control conditions were not shown the graph, but simply told that the expert team had estimated that costs between £90 million and £120 million were *likely* (Control 1) o*r most likely* (Control 2). They were subsequently asked to estimate the probability equivalents of these expressions, in the same way as above.

3.2. Results

How wide is a "likely" and a "most likely" interval? Participants in the Experimental conditions suggested on average that the expert team would describe costs between 83.4 million (SD = 19.0) and 128.5 million (SD = 20.3) as "likely", and costs between 87.8 million (SD = 17.9) and 125.4 million (SD = 16.4) as "most likely". None of the differences between these two conditions were significant. The mean estimated range of "likely" costs was 45.0 million (SD = 33.5), only slightly wider than the estimated range of "most likely" costs, 37.6 million (SD = 29.4). This difference did not reach significance with a two-tailed test, $t(226) = 1.789$, $p = .075$, but indicates that some participants had assumed that "most likely" could be used to describe a narrow interval corresponding to the highest point of the of the curve. In fact, the middle, narrow interval of 100–110 million was chosen by 33% of participants in the Most likely condition, but only by 20% of participants in the Likely condition. This answer was particularly frequent in the highly educated group.

How probable is a "likely" and a "most likely" estimate? The suggested numeric probabilities of *likely* and *most likely* were also quite similar, with mean estimates around 70% in the experimental conditions and slightly higher in the control conditions (see middle panel of Table 1). A 2 × 2 ANOVA with group (experimental vs. control) and phrase (likely vs. most likely) as the two independent factors revealed a significant difference for group, $F(1, 461) = 14.76$, $p < .001$, but none for phrase, $F(1, 461) = 1.01$, $p = .316$, and no significant interaction $F(1, 461) = 2.08$, $p = .151$. Probability estimates were approximately the same for participants with higher vs. lower levels of education.

Interestingly, the suggested probabilities were not related to interval width. Fig. 2 shows probability estimates of selected narrow, middle sized, and wide ranges. These three range values were the most common ones, corresponding to one, three, or five central bars, and containing together about 60% of all answers. The subjective probability estimates did not reflect the actual proportions of the distribution, but remained the same, with no apparent reduction for narrow ranges. In fact, the correlations between range and probability estimates were close to zero, with $r = -0.06$ and $r = -0.14$ for "likely" and "most likely", respectively. This indicates that participants' probability estimates were not derived from inspection of the corresponding areas under the curve.

Which intervals are likely? Despite their variability in width, almost all intervals are centrally located. The middle value in the distribution, £105 million, is in fact included in 98.2% of all proposed *likely* and *most likely* intervals. So, in this sense, both wide and narrow intervals are representative of the parent population and may be perceived as equally probable based on representativeness as the primary criterion.

3.3. Discussion

Participants in this study received a probability distribution binned into 13 smaller sections visualized as bars. This was supposed to make it easier to discriminate between a narrow "most likely" value and a larger "likely" interval. Yet very few respondents appeared to have distinguished between these two phrases. The partitioning was also intended to facilitate more realistic probability estimates. But the participants seemed not to use areas under the curve as a cue to probability, and gave on average the same subjective probabilities for narrow as for wider intervals. As a result, only respondents who suggested very wide intervals (80–130 million) gave estimates that corresponded roughly to the actual proportions of the probability distribution, as displayed in the graph.

Participants might have failed to take the areas under the curve into account because they did not see their relevance, but also because their

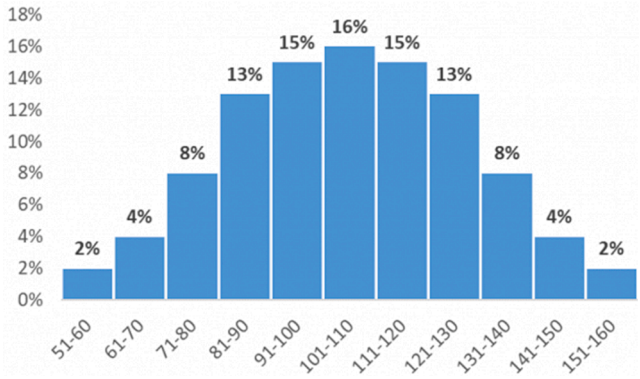


Fig. 3. Graph illustrating a stepwise probability distribution of estimated costs, with percentages appended, as used in Study 3.

magnitudes were not explicitly stated other than in a graphic format. As the y-axis was not labelled, and there were no numbers or percentages attached to the bars, the proportions could only be estimated by eyeballing the graph. In the next study, such numbers were added to make area information more salient and facilitate more accurate numerical estimates.

4. Study 3

In this study the objective proportions were made more explicit and accessible by appending the actual numeric percentages to the bars. Participants were randomly allocated to one of two preregistered studies (AsPredicted reference #59788). Study 3A was a replication of Study 2 with the new graph and included in addition a measure of graph literacy. Study 3B used a within-subjects design, where the same participants performed both the "likely" and the "most likely" judgment tasks. This kind of design is expected to alert participants to potential differences between rated objects (Birnbaum, 1999; Charness et al., 2012). We expected accordingly that estimates of "likely" and "most likely" intervals would differ more in Study 3B than in Study 3A, both with regard to interval magnitudes and to probability estimates (with lower probabilities for "most likely" estimates).

4.1. Method

Participants. Participants were recruited from the UK and Ireland via Prolific. They received the likelihood questions appended to an unrelated questionnaire on vaccination intentions. After excluding 14 participants who gave incomplete or ambiguous answers, the final sample consisted of 440 participants (304 women, 131 men, 5 other), with ages ranging from 17 to 71 years ($M = 34.6$, $SD = 12.4$). About half (53.2%) reported to have obtained a bachelor's degree or more. They were randomly assigned to one of three conditions, two between-subjects (Exp. 3A, $n = 146$ and 147 for the *Likely* and *Most likely* condition, respectively) and one with a within-subjects design (Exp. 3B, $n = 147$).

Material and procedure. All participants were given questionnaires with the same scenario as in the previous studies, accompanied by a slightly edited probability distribution of costs, displayed in Fig. 3. The main difference from Study 2 was a more compact distribution of costs, extending from 51 to 160 million (vs. 40–170 million in Study 2) distributed over 11 (rather than 13) bars, and included written percentages attached to the bars. In the between-subjects conditions (3A) participants were asked either to define a "likely" or a "most likely" interval, as before. In the within-subjects condition (3B) they were told about one expert who described an interval for "likely" costs, and another expert who described the "most likely" costs. The order of these questions was balanced across participants. Both orders yielded similar estimates, so data were pooled. Participants in all conditions were subsequently asked to estimate the numerical probability of costs associated with their proposed intervals. Participants in Study 3A finally completed a four-item short graph literacy measure (Okan et al., 2019), which yields scores ranging from 0 to 4.

4.2. Results of Study 3A

Participants proposed intervals of all sizes, ranging from narrow (101–110 million) to very wide (51–160 million), both in the *Likely* and the *Most likely* condition, with a mean width of about 35 million (see lower panel in Table 1). This is slightly less than in Study 2, probably because of the more compact shape of the distribution. The intervals suggested to describe "most likely" costs were in Study 3A indistinguishable from the intervals that were just "likely".

The probabilities attached to the selected intervals were also quite similar, with *most likely* slightly less probable than *likely*, but not significantly so; $t(291) = 1.80$, $p = .072$.

The bars in the graph were in this study headed by percentages, highlighting their relative areas under the curve. Thus, the probabilities associated with each interval could in this study easily have been derived from the presented numbers. For instance, costs of 101–110 million were 16% likely and costs of 91–120 million were 46% likely. But most participants did not make use of the provided numbers. Of 43 participants in the *Likely* condition who selected the middle narrow intervals, only ten suggested probabilities of 15–16%, whereas the majority proposed probabilities above 50%. In the *Most likely* condition, the match was better, but half of those who had selected the middle, narrow interval reported probabilities above 50% here, too.[1] For intervals larger than 20 million, the probabilities were on average closer to the actual areas under the curve, however, they did not seem to have been derived from the provided numbers in a consistent way. For instance, ranges of 30 million (matching the three central bars of the graph) were common, but only 11 out of 47 suggested probabilities that matched the percentages (of about 45%) in the graph, the others suggested lower or higher estimates.

Proposed interval magnitudes and probability estimates were positively, but weakly correlated; $r = 0.205$ ($p = .013$) and $r = 0.194$ ($p = .019$) for *likely* and *most likely*, respectively. This indicates that outcomes within wide intervals are estimated to be only slightly more likely to occur than outcomes from narrow intervals. The graph and the numbers provided should have suggested a much closer relationship.

Graph literacy: Mean number of correct answers on the graph literacy test = 2.27 ($SD = 0.82$). This is comparable to the mean score observed in a representative US sample: 2.22 ($SD = 1.10$) (Okan et al., 2019).

Judgments of ranges and their probabilities might presuppose ability to read and understand a graphical display, like the one shown in Fig. 3. We hypothesized that individuals with high graph literacy would propose wider ranges for "likely" costs, and smaller ranges for "most likely" ones, and also that they would suggest probabilities that better matched the ranges they had selected. But graph literacy scores were essentially unrelated to width of range, $r = 0.025$ for "likely", and $r = 0.027$ for "most likely" estimates. Graph literacy was weakly positively correlated with subjective probability estimates in the *Most likely* condition, $r = 0.19$, $p = .019$, but not in the *Likely* condition, $r = 0.05$, *ns*. There were no significant differences in probability estimates related to high vs. low levels of education.

4.3. Results of Study 3B

Participants in this within-subjects study were asked to generate both a "likely" and a "most likely" interval. The opportunity to compare these

[1] Similar exaggerated probability estimates were reported by Juanchich et al. (2022) for freely selected outcomes in a frequency distribution.

two phrases led a majority to propose intervals of different size, with 59% suggesting wider estimates for "likely", whereas 20% suggested that "most likely" costs were wider (about 20% estimated them to be the same). This implied a smaller mean interval for most likely, as displayed in Table 3, the difference between *likely* and *most likely* being highly significant with a paired-samples test, $t(146) = 4.54$, $p < .001$, Cohen's d = 0.38. Participants with a high level of education suggested somewhat wider ranges for both likely and most likely (Table D1 in Appendix D).

The probabilities attached to these "likely" and "most likely" intervals also differed, but not much: "Likely" intervals were judged on average 3% more probable. A paired samples test of the means reported in Table 3 yielded a significant difference, $t(146) = 2.22$, $p = .028$, Cohen's d = 0.18. For each term, the correlations between magnitude of proposed interval and estimated probability were small but positive, $r(147) = 0.14$, $p = .09$, and $r(147) = 0.18$, $p = .028$, for "likely" and "most likely", respectively. As in Study 3A, only a minority of those proposing the narrow middle category seem to have used the information about proportions (15–16%) as a guide to probability. Most probability estimates were 50% or higher even for this narrow set of outcomes.

4.4. Discussion

The three reported studies examined what is meant by *likely* in a context of projected costs. The studies led to several surprising findings. First, participants did not discriminate between "likely" and "most likely" (except when the same participants were required to judge both terms, and even then, they were not entirely consistent). Second, their subjective probability estimates were not informed by a graph showing the complete probability distribution; and third, attempts to make the task more transparent by dividing up the distribution in smaller bins, with explicit numeric information about their likelihood, had almost no effect on estimates. Third, the magnitude of a "likely" (or "most likely") range of costs appeared unrelated to their judged probabilities. This all suggests that people did not use "likely" in a normative, probabilistic sense, but preferred a pragmatic notion based on typicality. As in Study 2, nearly all participants in Study 3 (95.2% and 93.1% in the two conditions, respectively) proposed *likely* intervals that comprised £105 million, the middle value of the distribution. But since these values were also the most frequent ones, the selected "likely" outcomes could reflect both centrality and frequencies.

In the next three studies, we sought to unconfound these features, by asking people to judge central compared to peripheral intervals that are equally or more frequent. Finally, we asked in Study 6 for likely outcomes in skewed distributions where two measures of centrality, median and mode, did not coincide.

5. Study 4a

The ranges generated in the first three studies almost always spanned the middle values of the distribution, so another set of studies was added to examine which segments in a distribution, those in the center or in the tails, were perceived as *likely*, in contrast to outcomes that were deemed to be *not likely*. Again, a probabilistic definition of likely as $p > .6$ requires an outcome interval covering the greater part of the curve, whereas smaller segments, regardless of their location (in the tails or in the center of the distribution) should by this criterion be regarded as *not likely*. But people who use "likely" according to a pragmatic definition might base their selections on judgments by and of *representativeness* (Tversky & Kahneman, 1982): Peak and central outcome segments are arguably more representative of the distribution than peripheral ones and might according to a "representativeness heuristic" (Kahneman & Tversky, 1972, 1973) be viewed as likely more often than outcome segments situated in the tails.

Participants in Study 4 were asked which ones of several segments in the distribution they naturally would describe as being *likely* or *not likely*. If centrality is the main criterion, both central intervals should be described as "likely" regardless of their width, whereas comparable tail intervals would be "not likely".

We also investigated whether centrality affected participants' numerical probability assessments, or whether these estimates would correspond more closely to the objective proportions displayed in the graph.

The study was preregistered (AsPredicted reference #61223).

5.1. Method

Participants. Participants were recruited from Prolific. They received the questions after an unrelated task. After discarding six participants who did not comply with the instructions, or failed a control question, 187 questionnaires were retained for analysis, from 128 women and 59 men, 18–70 years old, mean age = 36.0 ($SD = 12.5$); half of them (55.2%) reported having obtained a bachelor's degree or higher. They were randomly allocated either to a verbal condition where they evaluated whether outcomes in selected intervals were *likely* or *not likely,* or to a numerical condition where they estimated the numerical probabilities of the same intervals.

Material and procedure. All participants in this study received the graph in Fig. 1 (showing intervals without explicit percentages), with the same scenario as in previous studies.

Verbal condition. Participants in this condition were asked: Would you characterize the following costs as likely or not likely? For each cost, select the expression that seems most right. (Order of statements was randomized between subjects.).

(a) Costs of less than 80 million: Likely or not likely?
(b) Costs between 90 and 120 million: Likely or not likely?
(c) Costs between 100 and 110 million: Likely or not likely?
(d) Costs of more than 130 million: Likely or not likely?

Intervals (a) and (d) focused on the tails and intervals (b) and (c) on the central values, with interval (b) being wider than (c). Costs in the ranges (a) (c) and (d) were approximately equally probable according to the graph (17.3%, 15.0%, and 18.0%). The wide central range (b) comprised a larger part (about 42%) of the distribution.

Table 2
Percentages of participants who described outcomes in the center or in the tail of the distribution as "likely" or "not likely" (Verbal condition; most frequent response in bold), along with numeric probability estimates of the same intervals (Numerical condition); Study 4a.

	Costs	Verbal condition % of answers		Numerical condition Mean estimates (SD)	
Outcome	Intervals	Likely	Not likely	Subjective probabilities	Objective percentages
(a) Lower tail	<80 mill.	24.0	**76.0**	30.6 (16.6)	17.3
(b) Wide central	90–120 mill.	**94.8**	5.2	56.5 (22.8)	42.1
(c) Narrow central	100–110 mill.	**87.5**	12.5	55.0 (27.0)	15.0
(d) Upper tail	>130 mill.	28.1	**71.9**	36.7 (19.4)	18.0

Table 3
Percentages of participants who describe outcomes in the center or in the tails of the distribution as "probable" or "not probable"; Study 4b.

Outcome	Interval	Probable	Not probable	Objective percentages, based on graph
(a) Lower tail	<90 million	44.4%	55.6%	28.6%
(b) Narrow central	100–110 million	75.0%	25.0%	15.0%
(c) Upper tail	>120 million	72.2%	27.8%	29.3%

Numerical condition. Participants in this condition received questions about the same intervals, to be answered with numeric probabilities. "How likely are these costs, in your opinion? Complete each statement with a probability between 0 and 100% that feels most right".

5.2. Results and discussion

If the term "likely" should be reserved for probabilities larger than 50% (typically requiring a 60–80% chance), as prescribed by most translation standards, *none* of these cost predictions should be judged as "likely". If "likely" signifies a representative, central outcome, statement (b) and (c) describe likely costs, whereas (a) and (d) do not. Inspection of individual response patterns revealed that that not a single participant conformed to the translation standards and judged all four outcomes as "not likely". The most frequent response pattern (given by 57.3% of the respondents) was "likely" for the two central intervals and "not likely" for the left and right tail intervals, as predicted by the centrality hypothesis. Altogether, around 90% of participants stated that the central intervals, both the wide one and the narrow one, were "likely", whereas costs in the tails of the distribution (less than 80 mill or more than 130 mill were judged "not likely" by a majority of about 75%, as shown in Table 2.

The right panel of Table 2 shows that the numeric probabilities of middle costs were estimated to be of the same magnitude both for wide and narrow intervals, with both estimates above 50% on average. Probability estimates for the tails were substantially lower, but still inflated compared to the objective percentages (as derived from the corresponding portions of the graph). If we consider all estimated p values above 20% in statements (a), (c), and (d) as over-estimates, we find that 70–80% of all probabilities were exaggerated, often to a large degree. The wide interval in (b) that comprised 42.1% of the distribution, was also overestimated, but not as much.

Both verbal and numeric estimates of "likely" were affected by their relative positions in the distribution (central vs. peripheral), rather than by the proportions displayed, confirming our hypothesis that the pragmatic usage of this term differs from its formal probabilistic meaning.

6. Study 4b

Participants in study 4a claimed than costs in the middle of a probability distribution could be characterized as "likely", even for a narrow middle segment, whereas outcomes from comparable regions in the tails were "not likely". The present study was set up to explore whether this difference between the centre and the tails still holds when the tail regions are expanded so as to contain a larger proportion than the central segment.

6.1. Method

Participants. Participants in this study were 72 first-year psychology students attending an online psychology lecture at a Norwegian university. Demographic data were not collected, but previous surveys indicate that these lectures are attended by a majority (about 75%) of women with a median age of 21 years. Most of these students also followed a course in introductory statistics during the same term.

Materials and procedure. They were shown the same graph as in 4a (displayed in Fig. 1) but asked to evaluate only three outcomes, namely (a) costs of less than 90 million, (b) between 100 and 110 million, and (c) more than 120 million. The tails in (a) and (c) contained each 28–29% of the distribution, and were accordingly nearly twice as likely as the narrow middle category in (b). For each of these outcomes they should indicate which expression, "probable" or "not probable"[2] that "feels more right".

6.2. Results and discussion

We replicated results from the verbal condition in 4a. Despite low formal probabilities for all described outcomes, the central outcome was described by a majority as "probable", as reported in Table 3. Also outcomes in the upper tail were considered more "probable" than not. In fact, the two most common patterns of answers were "probable" for all three outcomes (n = 18), and "probable" for the middle outcome and "not probable" for the tail outcomes, (n = 17). The preference for "probable" for outcomes in the upper tail may be due to popular beliefs about frequent overruns of large public projects (Flyvbjerg, 2014).

This study showed that narrow middle intervals were described as "probable" by a Norwegian-speaking student sample, despite being compared to other sections of the distribution that were almost twice as probable. Nobody said that all of the four outcomes were *not* probable, even if they all had objective probabilities below 50%, corresponding more closely to outcomes that in translation studies have been considered "unlikely" rather than likely.

7. Study 5

The common theme in all the previous studies was the likely costs of a highway construction project. To explore the generality of these findings, Study 5 presented similar probability distributions of outcomes in two other domains: climate and health. The information was illustrated by the same graph, but with different units along the x-axis, which indicated either a rise of sea level (in cm) or expected rates of COVID-19 (per 10,000 inhabitants). The intervals to be compared in this study covered together an exhaustive range of outcomes, and contrasted the central part to the remaining peripheral parts. The study was preregistered (AsPredicted #63188).

7.1. Method

Participants. Participants were recruited from the UK via Prolific. They received the question about likelihood appended to an unrelated questionnaire. After excluding one participant who withdrew from the study, the final sample consisted of 369 participants (267 women, 99 men, 3 other), with ages ranging from 18 to 87 years (M = 36.6, SD = 14.0). About half of them (55.2%) had bachelor's degree or higher. They were randomly allocated to one of four conditions, according to a 2 × 2 design, with scenario (climate vs. health) and interval set (wide vs. narrow central interval) as the two factors.

Material and procedure. All participants received a graph similar to the one presented in Study 2 and 4, spanning the entire distribution from

[2] The Norwegian terms were «sannsynlig»/ "ikke sannsynlig». There is in Norwegian no distinction between likely and probable. "Sannsynlig" is used both as a technical and a more colloquial term.

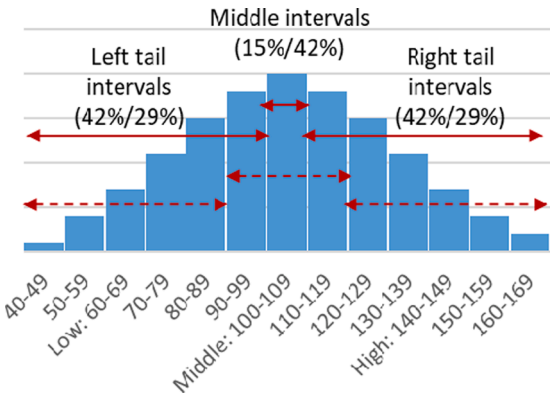


Fig. 4. Graph illustrating probability distribution of projected rise in sea level (in cm) or number of people (out of 10,000) contaminated by a new variant of COVID-19, for participants in Study 5. They were asked to evaluate whether the middle and the tail intervals could be described as "likely" in two conditions: Narrow central (solid lines) and Wider central (dashed lines). (Participants were shown the graph without the arrows.).

40 to 169 around a central, modal value of 100–109. These numbers now referred either to projected sea level rise (in cm) or to expected rate of infection of a new variant of COVID-19 (per 10,000 inhabitants), both according to expert estimates. The instructions had an added sentence explaining that the height of bars indicated probabilities, to make the task more transparent for lay participants (full materials are available in Appendix B).

Participants received a set of three "likely" statements, one referring to a central interval and the other two to the remaining lower and higher portions of the distribution. In the Narrow central condition (A), the central interval included only the middle, 100–109 bar, while the higher and lower intervals covered the remaining, more comprehensive portions of the distribution, as indicated by the solid arrows in Fig. 4. In the Wide central interval condition (B), this interval was spanning three bars (dashed arrows in the figure), with the tail areas reduced correspondingly. Participants in all conditions were asked to indicate their agreement with each statement on a Likert scale, from 1: Strongly disagree to 5: Strongly agree.

For the Sea Level the statements in Condition A [B] were (randomized order):

(a) The model indicates that a rise in sea level of less than 100 cm [B: 90 cm] is likely.
(b) The model indicates that a rise in sea level between 100 and 109 cm [B: 90–119 cm] is likely.
(c) The model indicates that a rise in sea level of more than 110 cm [B: 120 cm] is likely.

The middle interval in Condition A covered about 15% [B: 42%] of the distribution, and each of the peripheral intervals included about 42% [B: 29%]. Outcomes in the peripheral intervals were accordingly *more* probable than the middle interval in A, and *less* probable than the middle interval in Condition B.

Participants were subsequently asked to suggest numeric estimates of the probabilities for the same three intervals. To avoid their personal opinions on climate change and pandemics, they were asked third-person questions: How do you think the experts would estimate the probability of the three sea level rises [infection rates] below? Normatively, these estimates should reflect the corresponding areas in the graph, but they might also be affected by their centrality and the chosen verbal label ("likely" vs. "not likely").

7.2. Results and discussion

As shown in Table 4, participants in all conditions agreed that the middle interval should be described as "likely", both for intervals that were fairly large and covered 42% of the distribution, but also when the interval was narrow and comprised only 15%. The agreement scores in Table 4 show that peripheral outcomes were not considered "likely" even when they covered a greater area of the distribution than the central interval. Repeated measures analyses of agreement scores for statements show highly significant differences of means in all four conditions, $F(2, 90) = 27.98$, and $F(2, 90) = 25.28$ for Sea Rise and COVID-19 scenarios in the upper panel of Table 4, respectively, and $F(2, 92) = 73.68$ and $F(2, 88) = 49.60$ for the same scenarios in the lower panel of the table, all with $p < .0001$. In all these cases the middle interval stood out as being judged more likely than the two peripheral partitions. This pattern of answers could be found both for participants with high and low level of education, although highly educated participants gave a bit lower probability estimates in the COVID-19 vignette (see Table D3 in Appendix D).

An overall mixed 2 × 2 × 2 ANOVA of agreement scores with Centrality (central vs. mean of left and right tail) as a within-factor, and Scenario and Condition as between-factors, gave an overall effect of Centrality, $F(1, 363) = 241.63$, $p < .0001$, $\eta^2 = 0.40$, and of Scenario $F(1, 363) = 10.33$, $p < .001$, $\eta^2 = 0.028$, no main effect of Condition, but an interaction between Condition and Centrality, $F(1, 363) = 11.73$, $p < .001$, $\eta^2 = 0.031$, indicating that the effect of Centrality was largest in the Wide central interval condition, as predicted.

A comparison of the upper and lower panel in Table 4 confirms the importance of centrality. Outcomes in a 42% segment of the distribution are "likely" when the segment is a central one, as in the lower panel, but not in 42% segments located above or below the distribution midpoint, as in the upper panel.

Subjective probability estimates did not reflect the areas in the graph, even if the instructions explicitly stated that the height of bars indicated probabilities. In the Narrow central interval condition, mean estimates were similar for all intervals, regardless of the segment size, as shown in Table 4. Probabilities given for the central segment were accordingly grossly over-estimated. In the Wide central interval condition (lower panel), this interval was considered (correctly) more probable than the two peripheral segments; $F(2, 93) = 78.21$, $p < .0001$ and $F(2, 88) = 25.64$, $p < .0001$. However, the middle intervals were still overestimated compared to the objective proportions of these intervals. An overall mixed 2 × 2 × 2 ANOVA of probability estimates with Centrality (central vs. mean of left and right tail) as a within factor, and Scenario and Condition as between-factors, gave an overall effect of Centrality, $F(1, 365) = 60.61$, $p < .0001$, $\eta^2 = 0.14$ indicating that the central parts were judged more likely than the tails.

The three segments to be judged constituted an exhaustive, non-overlapping set of alternatives, whose probabilities should add up to 100%. But only about one fourth (26.3%) of the participants in the present study made estimates that could be considered additive by this criterion. This "additivity neglect" is in line with previous research that has shown that with multiple outcomes, most respondents violate the 100% convention for an exhaustive set of probabilities (Redelmeier et al., 1995; Riege & Teigen, 2013; Teigen, 1983), unless they are explicitly told to obey this rule. Such non-complementarity (or sub-additivity) is a robust finding for sets of multiple non-overlapping alternative outcomes (Sanbonmatsu et al, 1997; Van Wallendael & Hastie, 1990). The present results demonstrate non-complementarity for a continuous distribution partitioned in only three parts.

Agreement scores and probability estimates of corresponding segments of the distribution were similar for both scenarios, indicating that

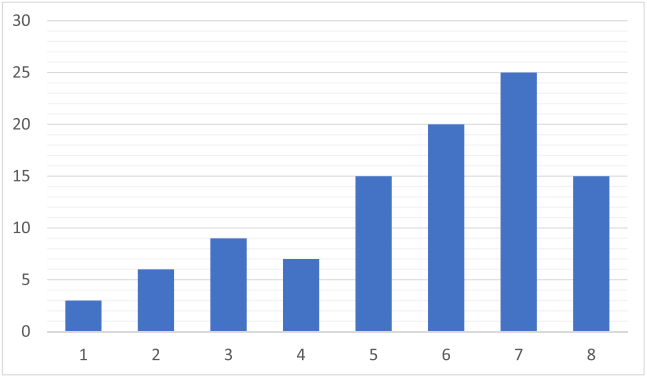


Fig. 5. Graph accompanying the questionnaire about likely student scores in the Continuous outcome vignette of Study 6. In the Categorical outcomes vignette, the scores 1–8 on the x-axis were replaced by country names (Canada, France, Germany, Italy, Japan, Russia, UK, and USA), in alphabetical order.

participants did not base their judgments on their own views of climate change or COVID-19 risks. The study demonstrates that the findings from the previous studies can be generalized to other domains and are not limited to judgments of project costs.

8. Study 6

All preceding studies were based on symmetric, bell-shaped distributions of continuous outcomes, where the middle, "likely" values were also the most frequent ones. Study 6 was conducted to extend the findings in three ways. First, we tested which outcomes participants would describe as "likely" in a skewed continuous outcome distribution, where the modal outcome did not occupy a central position, and differed from the median and mean. Second, we added a categorical outcome condition where the distribution included multiple discrete outcomes, listed in an arbitrary order. Third, we included control questions testing the participants' ability to read the distribution correctly. The study was preregistered (AsPredicted #7634).

8.1. Method

Participants. Participants were recruited from Prolific with quotas on gender, age and ethnicity to be representative of the UK population. After excluding 12 participants who failed the attention check and/or completed the study in less than one minute, the final sample consisted of 477 participants (246 women, 231 men, 0 other), with ages ranging from 18 to 81 years ($M = 44.6$, $SD = 15.2$). Of those reporting level of education, 59.5% had obtained a bachelor's degree or higher. They were randomly allocated two conditions, one describing a skewed distribution of quantitative (continuous) outcomes, the other a parallel set of discrete (categorical) events.

Material and procedure. All participants received a vignette describing an exam of a course in political science. Participants in the *Continuous outcome condition* were told that students would receive eight questions about the countries that formed the "Group of Eight" (G8) from 1997 to 2014, yielding a total score of up to 8 correct answers. A score distribution for 100 students was attached, as displayed in Fig. 5. They were then asked whether they would describe a score of 7, 6, and less than 5 (1, 2, 3, or 4) as "likely" or "not likely".

Participants in the *Categorical outcomes condition* were told that student at this exam would receive a question about one of these countries drawn from a test bank containing 100 questions varying from 3 (Canada) to 25 (about UK), as displayed in a categorical version of Fig. 5 (where the countries were listed alphabetically). They were then asked whether they would characterize an exam question about UK, Russia, or one of these four countries: Canada, Germany, France, or Italy, as "likely" or "not likely". (For a full set of the questionnaires, see Appendix C.).

To ascertain whether they had inspected and understood the information in the graph, participants in the continuous [categorical] outcomes condition were asked three simple control questions:

- How many students got a score of 6? [How many questions in the test bank are about Russia?]
- How many students got a score of 1, 2, 3 or 4? [How many questions in the test bank are about Canada, Germany, France, or Italy]
- How many students in total took the test? [How many questions are there in total in the test bank described above?]

8.2. Results and discussion

As shown in Table 5, the most frequent outcome (a score of 7 / a question about the UK) was, as predicted, described by a majority as "likely" rather than "not likely", despite the fact that this outcome would only have a 25% chance to occur. The second most frequent outcome (a

Table 4
Mean agreement scores (1–5) with "likely" statements for central and peripheral intervals and mean probability estimates (0–100) of corresponding outcomes in two conditions, Study 5.

Outcome	Interval	Objective percentages	Agreement scores		Probability estimates	
			Sea rise scenario	COVID-19 scenario	Sea rise scenario	COVID-19 scenario
Narrow central interval condition						
(a) Tail	<100	42.1%	3.38 (1.15)	3.05 (0.95)	48.5 (21.4)	40.3 (20.9)
(b) Center	100–109	15.0%	**4.20** (1.07)	**3.98** (1.04)	49.7 (29.1)	42.0 (27.6)
(c) Tail	>110	42.9%	3.52 (1.07)	3.45 (0.83)	45.5 (19.1)	43.8 (21.9)
Wider central interval condition						
(a) Tail	<90	28.6%	3.35 (1.11)	2.88 (1.04)	47.2 (23.0)	37.8 (21.0)
(b) Center	90–119	42.1%	**4.47** (0.81)	**4.18** (1.04)	63.9 (23.2)	52.8 (25.0)
(c) Tail	>120	29.3%	3.24 (0.96)	3.21 (0.99)	44.3 (21.8)	39.9 (22.6)

Table 5
Percentages of participants who describe modal, median and tail values of a skewed distribution as "likely" vs. "not likely" in Study 6.

Type of outcome	Score / Country	Objective percentages	Continuous distribution		Categorical distribution	
			Likely	Not likely	Likely	Not likely
Modal	7 / UK	25%	**65.1**	34.9	**89.0**	11.0
Median	6 / Russia	20%	**72.2**	27.8	**78.8**	21.2
Tail values	1–4 / four countries*	25%	20.4	**79.6**	39.6	**60.4**

* Canada, France, Germany, and Italy.

score of 6 / a question about Russia), which had a 20% chance, was also considered "likely" by a large majority in both conditions. The group of four least frequent outcomes forming the left tail of the distribution (scores of 1–4 or questions about the first four countries) were, in contrast, judged to be "not likely" by a majority, even if an outcome from this set also had a 25% probability to occur.

These results replicate and extend the findings from the preceding studies. The modal value (highest bar) is considered "likely" even in a skewed distribution. In this distribution the mode differs from the median and mean. The median (a score of 6) was considered about as likely as the mode (a score of 7), and a majority of 62% considered both these scores as likely. Thus, the term "likely" can be considered applicable to several representative values in the same distribution.

The pattern of responses appears not to be due to a failure of understanding information contained in the graph. Most participants answered correctly the three distribution comprehension questions (96% could read the number of cases for one outcome, 72% could add up the number of cases for three outcomes and 87% recognised the total sample shown in the distribution). In fact, 65% of the respondents answered all three questions correctly ($n = 311$) and when we compare them with participants who had made at least one error ($n = 166$), we found no evidence of a difference. Participants who answered all the comprehension questions correctly, selected "likely" as often to qualify the modal and median outcomes, and "not likely" as often for the set of tail outcomes, χ^2 (1) = 0.85, p =.770, φ = 0.01, χ^2 (1) = 0.69, p =.406, φ = 0.04, and χ^2 (1) = 0.01, p =.945, φ less than 0.01.

We expected, and found, the same general pattern of responses in both the continuous and categorical conditions. But the conditions also differed. Interestingly, "likely" seemed to be particularly applicable for outcomes in the categorical condition, both with respect to the most frequent outcome, and for the "tail" group of four infrequent outcomes, respectively; two-tailed χ^2 (1, $N = 477$) = 38.19, p <.001, φ = -0.28, and χ^2 (1) = 20.79, p <.001, φ = -0.21. For the median (2nd most frequent) outcome (a score of 6 vs. a question about Russia) the difference between categorical and continuous outcomes was in the same direction but was not statistically significant χ^2 (1) = 2.818, p =.093, φ = -0.08. These differences between conditions were not predicted and are not easy to explain. In the categorical condition, it is a bit misleading to call the group of four infrequent outcomes a "tail", since the countries were ordered arbitrarily and not according to their distance from a central value. This could contribute to explain that almost 40% said that those outcomes were likely, almost twice more often than in the continuous condition. But participants in the categorical condition selected "likely" more often also for the modal outcomes (compared to the continuous outcome conditions). Perhaps we unintentionally had made the vignettes in the two conditions different in terms of randomness, by implying (in the categorical condition) that the test questions were "drawn" from a test bank, while the exam scores (in the continuous condition) are not the product of a lottery, but are assumed to reflect a student's knowledge and degree of preparation for the test. It may be easier to say that all outcomes of a lottery procedure are "likely" in the sense that none of them can be ruled out, whereas exam scores are causally determined.

9. General discussion

"Likely" is perhaps the most frequently used linguistic term in a discourse of probabilities and risks (Juanchich et al., 2022), and plays a prominent role in most prescriptive scales of how to express probabilities in words (European Food Safety Authority et al., 2019; ICD, 2015; Mastrandrea et al., 2010). The present studies show, however, that the term does not have a stable meaning but means something different when used about a quantity than about a dichotomous fact. A "likely" quantity on a continuous scale or a distribution with more than two categorical outcomes says actually very little about its probability at all. This fact seems to have escaped the attention of most previous investigators of verbal probability. To our knowledge, the "binary assumption", on which most translation studies rest, has never been acknowledged, and it has implicitly been assumed that standard scales apply equally well to any type of outcomes. It has been concluded that VPEs are "vague" (Andreadis et al., 2021; Budescu & Wallsten, 1995; Wintle et al., 2019), context dependent (Harris & Corner, 2011; Weber & Hilton, 1990), and that their meanings reflect individual lexicons (Dhami & Wallsten, 2005; Karelitz & Budescu, 2004), whereas the distinction between likely binary facts and likely numbers has not been explored.

We aimed to test whether people understand and use "likely" in a way consistent with its probabilistic interpretation (p = 60-80%) or if it is rather used in a pragmatic way and designate a representative value in the distribution. In three studies, we gave participants a "likely" statement and asked them to generate corresponding intervals. We found that they did not distinguish between a "likely" and the "most likely" outcome; both phrases were assumed to describe approximately equal segments of the distribution and were associated with similarly high numeric probabilities. In the three subsequent studies, we gave participants intervals and ask whether they were likely or not. They chose again the central (median or modal) segments, regardless of the proportions included in these central parts. Peripheral (non-central) segments covering similar or even larger proportions, were, in contrast, not considered likely. Participants also overestimated the numeric probabilities of the chosen segments. Attempts to make the task more transparent by presenting the distribution graphically, and providing explicit information about percentages, did not make a difference, and demonstrate the robustness of these effects. By neglecting this crucial information, our lay participants behaved as if proportions of a distribution were irrelevant for defining "likely" outcomes or assessing their numeric probability.

9.1. Theoretical implications

Our studies led to the surprising findings that "likely" quantities and categorical outcomes with more than two alternatives are often rather improbable (p ≤ 40%). It seems sufficient that they are likely in a relative sense, namely compared to other, less likely quantities. This must be something more than just an imprecise manner of speaking, where a speaker actually *means* "most likely", but drops the modiifier, and says "likely" for short. When participants in Study 3 estimated "likely" and "most likely" outcomes side by side, only a few took the hint and used the opportunity to distinguish between these concepts. Joint presentations have in other studies encouraged respondents to discriminate between the concepts to be judged, assuming that they must be different since they are asked two questions rather than just one (Schwarz, 1996). Moreover, the same proportion of the curve was called "likely" when located in the center and "not likely" when located in the tails.

Observations like these suggest that judgments of "likely" quantities are not based on proportions, but on an outcome's typicality, or how well it mirrors central features of the distribution from which it is drawn. This feature was in Kahneman and Tversky's heuristics and biases approach labelled representativeness (Kahneman & Frederick, 2002; Kahneman & Tversky, 1972, 1973; Tversky & Kahneman, 1982). Representativeness, or typicality, was made responsible for several cognitive biases (Griffin et al., 2012; Teigen, 2022) but has been criticised for conceptual vagueness as a one-word label that "at once explain too little and too much" (Gigerenzer et al., 1999, p. 28). In a context of quantities, it can be given a more precise definition as a central, summary characteristic.

Measures of centrality hold a special place in descriptive statistics, and have been defined as "a representative value around which the measurements are distributed" (Bhattacharyyia & Johnson, 1977, p. 27). The present results indicate that *likely* is indeed used to describe a representative value, although we could not tell from the unimodal and

symmetric distributions used in Study 1–5 which measure of centrality (mean, median, or mode) was more important. From the skewed distributions in Study 6 both mode and median were chosen as instances of "likely" outcomes.

When outcomes vary along a continuous dimension, the distinction between "likely" and "not likely" outcomes becomes itself an exercise of dichotomization, achieved by contrasting a set of typical or regular (central) outcomes with those that are uncommon and deviant, and retaining the first set as being "likely". With strict criteria for being admitted to this set, it could be much narrower than a range embracing 70% of all outcomes. A parallel can be drawn to studies of overconfidence (over-precision) in forecasting, where experts fail to incorporate outcomes that they do not expect within the boundaries of their confidence interval, and consequently produce interval estimates that are too narrow (Moore & Healy, 2008; Teigen & Jørgensen, 2005).

Yet, after choosing a set of outcomes that were (according to the graph) in many cases less than 50% probable, participants claimed that the "likely" sets they had selected had a probability of 60–70%. Thus, they retained the standard numerical translation of "likely" despite using this term to describe a much less frequent event. This indicates a double standard: One for choosing which outcomes "likely" describes and another for defining "likely" in terms of numerical probabilities.

The distinction between two meanings of "likely" is reminiscent of the debate in linguistics about a semantic-pragmatics divide. The lexical or "literal" semantic meanings of a term are not always identical to the more subtle or indirect ways it can be used in a conversational setting. Even numbers that denote, by definition, a specific countable or measurable amount, can pragmatically permit an inexact interpretation, such as a lower bound reading of an interval (Levinson, 2000). Thus, 200 lives saved out of 600 in the Asian Disease Problem (Tversky & Kahneman, 1981) are formally equivalent to 400 lives lost, but may pragmatically be different by indicating that *at least* 200 are saved (Mandel, 2014; Fisher, 2020). In this case fewer than 400 lives may be lost. Similarly, we find that the standard dictionary definition of "likely" is "having a high chance of occurring or being true" (Merriam-Webster, n. d.), whereas it is pragmatically used to include events that do not have a high chance, but appear plausible for other reasons. This fits with the finding from Study 6 that more outcomes appear likely when drawn by a chance process than when determined by more stable causes.

The double standard we find for "likely" in these studies is paralleled by similar discrepancies for other VPEs in previous research. For instance, "possible" and "can" are commonly used to denote top outcomes in a distribution, and yet translated into numeric probabilities as high as 50%, by the same participants (Teigen et al., 2018). We can only speculate about the reasons for such puzzling discrepancies. In the case of "likely" it could be a consequence of conflating continuous with dichotomous distributions, so when people are asked to produce a probability estimate, they spontaneously perform an act of dichotomization, and start to think and make probability assessments as if it were a binary issue. They change, in Fox and Rottenstreich's (2003) terminology, from a class-based to a case-based approach to uncertainties, reducing the number of alternative outcomes from several to only two. When several such judgments are performed, they may add up to a total probability of more than 100%, as suggested by the inflated sums in Study 5. Alternatively, people make separate estimates of probabilities of an "aleatory" and an "epistemic" kind (Hacking, 1975), reflecting external vs. internal sources of uncertainty (Kahneman & Tversky, 1982). These are often referred to by different terms (Ülkümen et al., 2016), like *chances* vs. degrees of *confidence*. Participants in the present studies may have viewed the graphs as depicting external, objective probabilities, whereas the estimates attributed to "experts" were of a personal, internal kind, which might be assessed in another way and do not have to be identical to objective probabilities (Løhre & Teigen, 2016). A third, related possibility, is that people try to estimate the experts' second-order probabilities (their certainty about the stated probabilities), or a combination of the two (see Herbstritt & Franke, 2019, for a model of such combinations).

It may be tempting to describe the anomalous uses of "likely" revealed in the present studies as another instance of a probabilistic fallacy or cognitive bias. In that case it might be less common among highly educated people or those with a special background in statistics. However, we did not find consistent evidence of a relationship between education and use of terms in our studies, and no correlation between range estimates and graph literacy (Study 3A). Thus, we do not have to conclude that people are ignorant about the formal, statistical meaning of numerical probabilities, but rather that it is being overruled by their common understanding of the pragmatics of words.

9.2. Implications for uncertainty communication

Consumers of probabilistic messages should be aware that "likely" amounts mean something different from "likely" dichotomous facts. If a weather forecaster announces that *rain tomorrow is likely*, the chances of rain may indeed be 70% (with a complementary 30% chance of no rain). But if the forecaster says that *5 mm of rain is likely*, the probability of this amount is unspecified – the forecaster may only mean that 5 mm is a representative outcome in a distribution of expected amounts. For intervals the situation is even more ambiguous: If 4–6 mm is "likely" it *could* mean a 70% chance for an amount of rain within this interval, or it could simply indicate the forecaster's best guess, especially in the case of narrow ranges.

The present findings suggest that standard interpretations and standard scales of verbal terms should come with a caveat: The translations offered are only valid for dichotomous events. The guidelines of IPCC (International Panel of Climate Change, Mastrandrea et al., 2010) reveal a binary assumption by describing probabilities from 33% to 66% as "About as likely as not".[3] The NATO standard (North Atlantic Treaty Organization, 2016) explains middle probabilities of 40%-60% as "even chance", implying that just two options are compared. The implicit assumption of binary complementarity cannot be transposed to quantities or to situations with more than two competing outcomes (e.g., what will be the sea level in the future, who will win the tournament, or be selected for a job). In all the present studies, the outcomes people described as "likely" were actually *less* likely to occur than not, and should technically speaking be described as "not likely".

Experts with full knowledge of the distribution may be aware of the status of their verbal phrases, so if a project is "likely" to cost £100-110 million, the speaker knows whether the phrase refers to a dichotomy, namely the likelihood of costs *within* vs. *outside* of this interval, or just indicates a central and representative outcome interval from which no inference about a specific probability can be drawn (and hence implying no advice about how much to bet). However, for a receiver without this background information, "likely" is ambiguous, and could lead to a too strong reliance on a specific estimate, or (perhaps more worrisome) a failure to prepare a "plan B" in case a "not likely" tail event should occur, which could in fact be equally or more probable.

Dichotomies are often used to encourage and justify decisions (DeCoster et al., 2009). Policymakers, who receive and perhaps base their decisions on verbal probability expressions, may be misled to believe that a "likely" scenario is one that is expected to occur. A "likely" threat calls for preventive measures, whereas one that is "not likely" may be neglected as not requiring immediate action. It is accordingly not just an academic question which outcomes should be labelled "likely".

[3] And yet the IPCC report occasionally use "likely" to qualify numeric quantities. E.g., "Global average sea level in the last interglacial period (about 125,000 years ago) was likely 4 to 6 m higher than during the 20th century" (Quoted in Budescu et al., 2009, p. 3).

10. Concluding remarks

In their interdisciplinary overview of research on uncertainty communication, van der Bles et al. (2019) distinguished three objects to be uncertain about, namely facts, numbers, and hypotheses (models). All these uncertainties can be expressed in different formats (verbal, numerical, or visual). Our work adds an extra layer to this model by showing that format effects are not independent of object, specifically that verbal expressions mean something different when they are applied to uncertain quantities (numbers) than when they are applied to uncertain facts (dichotomous events).

The usage of "likely" in the present study indicates that this term reflects, for quantities, outcome representativeness rather than outcome frequencies. We believe that these findings also extend to the near-equivalent term "probable", which for most purposes can be used interchangeably with "likely", and is in many languages translated with the same term (Doupnik & Richter, 2002 - see also Study 4b). Further studies where "likely" and "probable" are combined with modifiers or prefixes ("very likely", "improbable") should be conducted to test the generality of these findings across related concepts. Another approach would be to offer alternative VPEs (possible, uncertain) and ask if any of these are more appropriate than "likely" to characterize central, but not highly probable values. A step in this direction was provided in Study 4 and 6, where "not likely" was used as an alternative response. This term was considered appropriate only for describing outcomes in the tails of the distribution. Finally, we might test the representativeness interpretation of "likely" by asking for descriptions of a wider interval that includes both central and more deviant outcomes. For instance, in Study 4, the 100–110 million interval was considered likely. Would larger, perhaps skewed intervals that include both central and more peripheral values, e.g., 70–110 million, be described as *more* likely (because of greater scope) or *less* likely (because of its inclusion of non-representative events)?

There is a long and still active research tradition devoted to converting verbal probability expressions into numerical equivalents. The general conclusion of this endeavor is that verbal phrases are generally vague and "fuzzy", so they may describe a subset of the 0–1 probability dimension rather than an exact value (Dhami & Mandel, 2022). In line with this, both experts and lay people claim that "likely" describe probabilities in the 60–80% range. Yet participants in the present studies agreed that outcomes with a probability as low as 15% could be "likely". This cannot be explained by the inherent vagueness of verbal terms, but rather by their non-probabilistic meanings. Defining the probability of "likely" is not possible without assumptions about the nature of the outcome and its distribution.

CRediT authorship contribution statement

Karl Halvor Teigen: Conceptualization, Writing – review & editing, Investigation, Formal analysis. **Marie Juanchich:** Investigation, Methodology, Visualization. **Erik Løhre:** Investigation, Methodology.

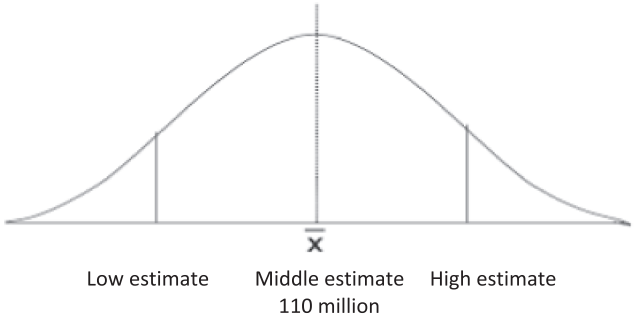
Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A:. Questionnaires for Study 1

Uncertain costs

The costs of major public investment projects (construction of roads, schools, hospitals) are being subjected to thorough quality assurance prior to start-up of construction. Despite detailed planning, there will always be considerable uncertainty involved. Thus, it is common that reports presented to decision makers comprise an uncertainty analysis stating the expected costs as a probability distribution like the one showed graphically in the figure below. This graph includes a middle estimate along with a low and a high estimate (called minimum and maximum), where 15 percent is below the low value and 15 percent is above the high value.



.

In a road sector project in Eastern Norway, a team of experts estimated an expected cost of £110 million (NOK 1100 million) for the recommended alternative, as shown in the graph.

Condition 1

Imagine that the expert team is asked to make a statement about the ***likely*** costs of this road project. What do you think they will answer (fill in the slots)

Likely costs will be between ……. million and ……. million

The experts are then asked to state the probability of these "likely" costs, on a scale from 0 to 100 percent. What do you think they will answer?

By the expression "likely" costs we mean costs that have a probability of about …….. %

Condition 2

Imagine that the expert team is asked to make a statement about the ***most likely*** costs of this road project. What do you think they will answer (fill in the slots)
Most likely costs will be between ……. million and ……. million
The experts are then asked to state the probability of these "most likely" costs, on a scale from 0 to 100 percent likely. What do you think they will answer?
By the expression "most likely" costs we mean costs that have a probability of about …….. %

Control condition 1

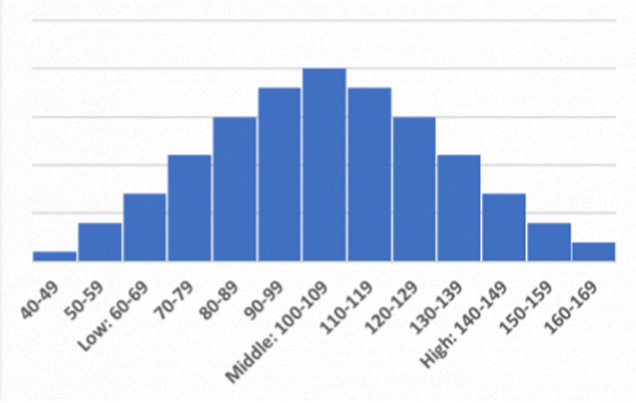
Imagine that an expert team is asked to estimate expected costs for a road project. They say that costs between £100 million and £120 million are ***likely.***
If they were asked to state the probability of these "likely" costs on a scale from 0 to 100%, what do you think they will answer?
By the expression "likely" costs we mean costs that have a probability of about …….. %

Control condition 2

Imagine that an expert team is asked to estimate expected costs for a road project. They say that costs between £100 million and £120 million are ***most likely.***
If they were asked to state the probability of these "most likely" costs on a scale from 0 to 100%, what do you think they will answer?
By "most likely" costs we mean costs that have a probability of about …….. %

Appendix B:. Questionnaires for Study 5

Sea level vignette



.

Climate scientists expect a rise in ocean level towards the end of the present century, as a result of global warming, but it is difficult to predict the exact magnitude of that rise.

A group of climate experts has calculated the probability of several scenarios for the rise in sea level around the coast of Iceland by the year 2100, as illustrated by the graph below. The horizontal axis shows the magnitude of the sea level rise in cm, and the height of the bars is an indicator of the likelihood of different rise magnitudes.

The graph indicates that a rise in sea level of less than 100 cm [B: 90 cm] is likely.
Strongly agree (1) (2) (3) (4) (5) Strongly disagree.
The graph indicates that a rise in sea level of between 100 and 109 cm [B: 90 and 119 cm] is likely.
Strongly agree (1) (2) (3) (4) (5) Strongly disagree.
The graph indicates that a rise in sea level of more than 110 cm [B: 120 cm] is likely.
Strongly agree (1) (2) (3) (4) (5) Strongly disagree.
How do you think the experts would estimate the probability of the three sea level rises below?

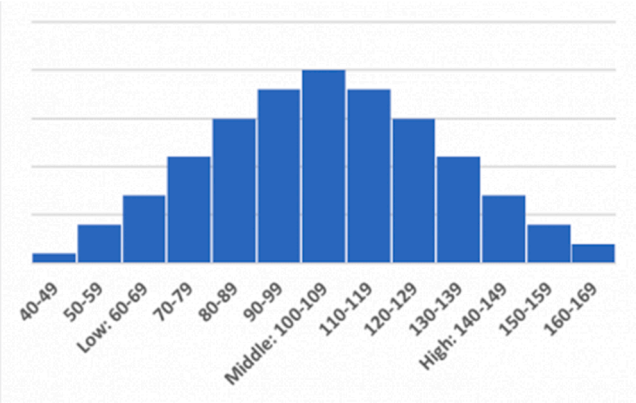
Complete each statement with a probability between 0% and 100% by adding your answer in the space provided.

- A sea level rise of less than 100 cm [90 cm] is …% likely.
- A sea level rise between 100 and 109 cm [90 and 119 cm] is … % likely.
- A sea level rise of more than 110 cm [120 cm] is … % likely.

COVID-19 vignette

The health authorities of a European country have modelled the number of people who could be infected by a new COVID-19 variant which causes severe form of the illness.

The medical experts made a model of the rate of infection which shows the number of individuals who could contract this new variant out of 10,000 inhabitants, incorporating the uncertainties involved. This model is shown in the graph below where the horizontal axis shows the infection rates (i.e., number of infection cases per 10,000 inhabitants) and the height of the bars shows the likelihood for these different infection rates.



.

The graph indicates that a rate of infection of less than 100 [B: 90] cases (per 10,000) is likely.
Strongly agree (1) (2) (3) (4) (5) Strongly disagree.
The graph indicates that a rate of infection between 100 and 109 [B: 90 and 119] is likely.
Strongly agree (1) (2) (3) (4) (5) Strongly disagree.
The graph indicates that a rate of infection of more than 110 [B: 120] (per 10,000) is likely.
Strongly agree (1) (2) (3) (4) (5) Strongly disagree.
How do you think the experts would estimate the probability of the three infection rates below?
Complete each statement with a probability between 0% and 100% by adding your answer in the space provided.

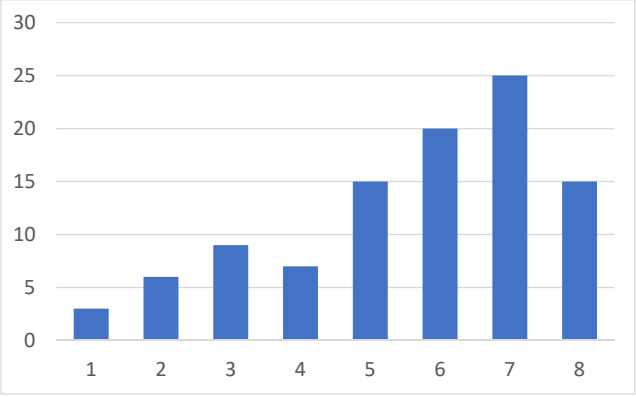
- A sea level rise of less than 100 [90] cases (per 10,000) is …% likely.
- A sea level rise between 100 and 109 [90 and 119] cases (per 10,000) is … % likely.
- A sea level rise of more than 110 [120] cases (per 10,000) is … % likely.

Appendix C:. Questionnaires for Study 6

Continuous outcome condition

G8

At the end of a course in political science, a student is told they will receive questions about the countries that formed the "Group of Eight" (G8) from 1997 to 2014 (Canada, France, Germany, Italy, Japan, Russia, United Kingdom, and USA). Each answer is scored as right or wrong, yielding a total score of correct answers that varies from 1 to 8. Of 100 students who have answered this exam, three students obtained a score of 1, 25 students obtained a score of 7, and so on, as displayed in the graph (where scores are listed from low to high).



.

Would you characterize the following possibilities as "likely" or "not likely"? For each sentence, select the expression that seems most appropriate.

	Likely?	Not likely?
The student will get a score of 7 on this exam.	☐	☐
The student will get a score of 6 on this exam.	☐	☐
The student will get a score of less than 5 (1, 2, 3 or 4) on this exam.	☐	☐

Control questions.

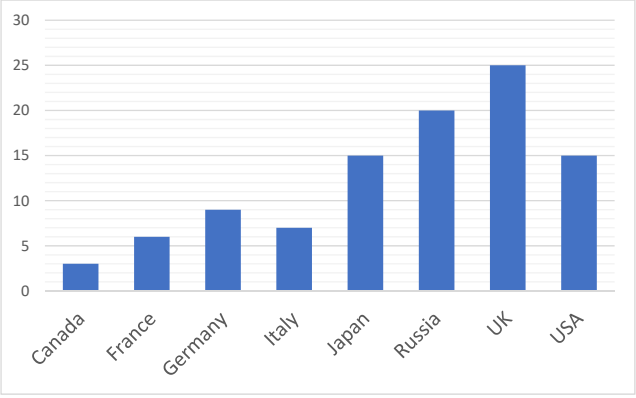
Based on the performance at the exam last year as described in the scenario and the graph above, could you please answer the following questions?

- How many students got a score of 6?
- How many students got a score of 1, 2, 3 or 4?
- How many students in total took the test?

Categorical outcomes condition

G8

At the end of a course in political science, students are told they will receive a question about one of the countries that formed the "Group of Eight" (G8) from 1997 to 2014. The question will be drawn from a test bank containing 100 questions. The students notice that the number of questions about each of these countries varies from 3 (Canada) to 25 (UK), as displayed in the graph (where the countries are listed alphabetically).



.

Would you characterize the following possibilities as "likely" or "not likely"? For each sentence, select the expression that seems most appropriate.

	Likely?	Not likely?
The exam question will be about UK	☐	☐
The exam question will be about Russia	☐	☐
The exam question will be about one of these four countries: Canada, Germany, France, or Italy	☐	☐

Control questions.

Based on the test bank described in the scenario and the graph above, could you please answer the following questions?

- How many questions in the test banks are about Russia?
- How many questions in the test bank are about these four countries altogether: Canada, Germany, France, and Italy?
- How many questions are there in total in the test bank described above?

Appendix D:. Effects of high vs low level of education

(See Table D1-D3).

Table D1
Mean ranges of Likely and Most likely intervals (in Millions of pounds) for participants with higher vs. lower levels of education (Studies 1–3).

Study	Condition	Education[1]		Effects of education[2]	
		Lower level	Higher level	*F* (high /low)	Significance
Study 1	Likely	44.4 M	45.1 M		
	Most likely	40.9 M	50.9 M	$F(1, 100) = 0.42$	$p = .52$
Study 2	Likely	39.8 M	47.8 M		
	Most likely	41.9 M	34.1 M	$F(1, 216) = 0.00$	$p = .99$
Study 3a Between-Ss	Likely	37.1 M	34.6 M		
	Most likely	32.1 M	38.5 M	$F(1, 280) = 0.36$	$p = .55$
Study 3b Within-Ss	Likely	29.7 M	38.7 M		
	Most likely	20.5 M	26.5 M	$F(1, 142) = 6.06$	$p = .015$

[1] Educational background were coded high vs low according to a median split, with Lower level corresponding to less than 4 years of academic education, and Higher level: bachelor degree or more (participants reporting "other" education were omitted from these analyses).
[2] There were no significant interaction effects.

Table D2
Mean subjective probability estimates of Likely and Most likely intervals for participants with higher vs. lower level of education (Studies 1–3).

Study	Condition	Education		Effects of education	
		Lower level	Higher level	F (high /low)	Significance
Study 1 Experimental	Likely	57.6%	65.2%		
	Most likely	62.7%	70.4%	$F(1, 102) = 3.13$	$p = .08$
Study 1 Control	Likely	73.1%	72.7%		
	Most likely	75.1%	75.1%	$F(1, 102) = 0.48$	$p = .49$
Study 2 Experimental	Likely	67.9%	66.5%		
	Most likely	65.5%	67.5%	$F(1, 216) = 0.13$	$p = .91$
Study 2 Control	Likely	68.6%	71.7%		
	Most likely	70.9%	76.7	$F(1, 226) = 3.55$	$p = .06$
Study 3a Between-Ss	Likely	52.3%	57.9%		
	Most likely	51.3%	46.2%	$F(1, 280) = 0.01$	$p = .94$
Study 3b Within-Ss	Likely	58.3%	56.4%		
	Most likely	55.8%	50.2%	$F(1, 142) = 0.74$	$p = .39$

Table D3
Mean subjective probability estimates of Narrow and Wide middle intervals for participants with higher vs. lower level of education for, Study 5.

Vignette/ Condition	Percentage of distribution	Low level of education	High level of education	*F*(high/low)	Significance
Sea level					
A: Narrow central (100–109 cm)	15.0%	46.8%	52.0%		
B Wide central (90–119 cm)	42.1%	66.3%	62.8%	$F(1, 175) = 0.04$	$p = .84$
COVID-19					
A: Narrow central (100–109 infected)	15.0%	48.4%	35.2%		
B: Wide central (90–119 infected)	42.1%	55.0%	52.1%	$F(1, 172) = 4.25$	$p = .04$

References

Andreadis, K., Chan, E., Park, M., et al. (2021). Imprecision and preferences in interpretation of verbal probabilities in health: A systematic review. *Journal of General Internal Medicine*. https://doi.org/10.1007/s11606-021-07050-7

Birnbaum, M. H. (1999). How to show that 9 > 221: Collect judgments in a between-subjects design. *Psychological Methods, 4*, 243–249.

Bhattacharyyia, G. K., & Johnson, R. A. (1977). *Statistical concepts and methods*. New York: Wiley.

Budescu, D. V., Broomell, S., & Por, H.-H. (2009). Improving communication of uncertainty in the reports of the Intergovernmental Panel on Climate change. *Psychological Science, 20*, 299–307.

Budescu, D. V., Por, H.-H., Broomell, S. B., & Smithson, M. (2014). The interpretation of IPCC probabilistic statements around the world. *Nature Climate Change, 4*(6), 508–512. https://doi.org/10.1038/nclimate2194

Budescu, D. V., & Wallsten, T. S. (1995). Processing linguistic probabilities: General principles and empirical evidence. *The Psychology of Learning and Motivation, 32*, 275–318.

Budescu, D. V., Weinberg, S., & Wallsten, T. S. (1988). Decisions based on numerically and verbally expressed uncertainties. *Journal of Experimental Psychology: Human Perception and Performance, 14*(2), 281–294. https://doi.org/10.1037/0096-1523.14.2.281

Clark, D. A. (1990). Verbal uncertainty expression: A critical review of two decades of research. *Current psychology: Research and reviews, 9*, 203–235.

Charness, G., Gneezy, U., & Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior & Organization, 81*, 1–8.

Collins, P. J., & Hahn, U. (2018). Communicating and reasoning with verbal probability expressions. *Psychology of Learning and Motivation, 69*, 67–105. https://doi.org/10.1016/bs.plm.2018.10.003

Collins, R. N., & Mandel, D. R. (2019). Cultivating credibility with probability words and numbers. *Judgment and Decision Making, 14*(6), 683–695.

DeCoster, J., Iselin, A. M. R., & Gallucci, M. (2009). A conceptual and empirical examination of justification for dichotomization. *Psychological Methods, 14*(4), 349–366.

Dhami, M. K., & Mandel, D. R. (2022). Communicating uncertainty using words and numbers. *Trends in Cognitive Science*. https://doi.org/10.1016/j.tics.2022.03.002

Dhami, M. K., & Wallsten, T. S. (2005). Interpersonal comparison of subjective probabilities: Toward translating linguistic probabilities. *Memory & Cognition, 33*, 1057–1068.

Doupnik, T. S., & Richter, M. (2002). Interpretation of uncertainty expressions: A cross-national study. *Accounting, Organizations and Society, 28*(1), 15–35.

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. https://doi.org/10.1016/j.obhdp.2010.10.005

Erev, I., & Cohen, B. L. (1990). Verbal versus numerical probabilities: Efficiency, biases, and the preference paradox. *Organizational Behavior and Human Decision Processes, 45*, 1–18. https://doi.org/10.1016/0749-5978(90)90002-Q

European Food Safety Authority. (2019). Guidance on Communication of Uncertainty in Scientific Assessments. *EFSA Journal*. https://doi.org/10.2903/j.efsa.2019.5520

Fisher, S. (2020). Meaning and framing: The semantic interpretations of psychological framing effects. *Inquiry*. https://doi.org/10.1080/0020174X.2020.1810115

Flyvbjerg, B. (2014). What you should know about megaprojects, and why: An overview. *Project Management Journal, 45*(2), 6–19.
Fox, C. R., & Rottenstreich, Y. (2003). Partition priming in judgment under uncertainty. *Psychological Science, 14*, 195–200.
Gigerenzer, G., Todd, P. M., & the ABC Research Group (1999). *Simple heuristics that make us smart*. Oxford: Oxford University Press.
Grice, H. P. (1975). Logic and conversation. In P. Cole, & J. L. Morgan (Eds.), *Syntax and semantics 3: Speech acts*. New York: Academic Press.
Griffin, D., Gonzalez, R., Koehler, D., & Gilovich, T. (2012). Judgmental heuristics: A historical overview. In K. Holyoak, & R. Morrison (Eds.), *Oxford Handbook of Thinking and Reasoning* (pp. 322–345). Oxford: Oxford University Press.
Hacking, I. (1975). *The emergence of probability: A philosophical study of early ideas about probability, induction and statistical inference* (2nd ed.). Cambridge: Cambridge University Press.
Harris, A. J. L., & Corner, A. (2011). Communicating environmental risks: Clarifying the severity effect in interpretations of verbal probability expressions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*, 1571–1578.
Herbstritt, M., & Franke, M. (2019). Complex probability expressions & higher-order uncertainty: Compositional semantics, probabilistic pragmatics & experimental data. *Cognition, 186*, 50–71.
Icd. (2015). Analytic Standards. *US Intelligence Community Directive, 203*. https://www.dni.gov/files/documents/ICD/ICD%20203%20Analytic%20Standards.pdf.
Irwin, D., & Mandel, D. R. (2020). *Variants of vague verbiage: Intelligence community methods for communicating probability. NATO STO technical report TR-SAS-114*. Toronto: Defence Research and Development Canada.
Jenkins, S., Harris, A. J. L., & Lark, R. M. (2018). Understanding 'unlikely (20% likelihood)' or '20% likelihood (unlikely)' outcomes: The robustness of the extremity effect. *Journal of Behavioral Decision Making, 31*(4), 572–586. https://doi.org/10.1002/bdm.2072
Juanchich, M., & Sirota, M. (2020). Do people really prefer verbal probabilities? *Psychological Research, 84*, 2325–2338.
Juanchich, M., Sirota, M., & Bonnefon, J.-F. (2019). Verbal uncertainty. In C. Cummins & N. Katso, *The Oxford Handbook of Experimental Semantics and Pragmatics*. doi: 10.1093/oxfordhb/9780198791768.013.2.
Juanchich, M., Sirota, M., & Teigen, K. H. (2022). People prefer to predict average and most likely outcomes, but over-estimate their likelihood. *Working paper*. University of Essex.
Juanchich, M., Teigen, K. H., & Gourdon, A. (2013). Top scores are possible, bottom scores are certain (and middle scores are not worth mentioning): A pragmatic view of verbal probabilities. *Judgment and Decision making, 8*(3), 345–364.
Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribution substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 49–81). Cambridge. UK: Cambridge University Press.
Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3*, 430–454.
Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80*, 237–251.
Kahneman, D., & Tversky, A. (1982). Variants of uncertainty. *Cognition, 11*, 143–157.
Karelitz, T. M., & Budescu, D. V. (2004). You say "probable" and I say "likely": Improving interpersonal communication with verbal probability phrases. *Journal of Experimental Psychology: Applied, 10*(1), 25–41. https://doi.org/10.1037/1076-898X.10.1.25
Levinson, S. C. (2000). *Presumptive meanings*. Cambridge, MA: The MIT Press.
Lichtenstein, S., & Newman, J. R. (1967). Empirical scaling of common verbal phrases associated with numerical probabilities. *Psychonomic Science, 9*, 563–564.
Lipkus, I. M., & Peters, E. (2009). Understanding the role of numeracy in health: Proposed theoretical framework and practical insights. *Health Education & Behavior, 36*(6), 1065–1081. https://doi.org/10.1177/1090198109341533
Lipkus, I. M., Samsa, G., & Rimer, B. K. (2001). General performance on a numeracy scale among highly educated samples. *Medical Decision Making, 21*, 37–44.
Løhre, E., Juanchich, M., Sirota, M., Teigen, K. H., & Shepherd, T. G. (2019). Climate scientists' wide prediction intervals may be more likely but are perceived to be less certain. *Weather, Climate and Society, 11*(3), 565–575. https://doi.org/10.1175/WCAS-D-18-0136.1
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*, 369–396.
Løhre, E., & Teigen, K. H. (2017). Probabilities associated with precise and vague forecasts. *Journal of Behavioral Decision Making, 30*, 1014–1026. https://doi.org/10.1002/bdm.2021
Mandel, D. R. (2014). Do framing effects reveal irrational choice? *Journal of Experimental Psychology: General, 143*(3), 1185–1198.
Mandel, D. R., & Irwin, D. (2021). Facilitating sender-receiver agreement in communicated probabilities: Is it best to use words, numbers or both? *Judgment and Decision Making, 16*(2), 363–393.
Mastrandrea, M. D., Field, C. B., Stocker, T. F., Edenhofer, O., Ebi, K. L., Frame, D. J., et al. (2010). *Guidance note for lead authors of the IPCC fifth assessment report on consistent treatment of uncertainties*. Intergovernmental Panel on Climate Change (IPCC).
Merriam-Webster. (n.d.). Likely. In *Merriam-Webster.com dictionary*. Retrieved November 20., 2021, from https://www.merriam-webster.com/dictionary/likely.
Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review, 115*(2), 502–517.
North Atlantic Treaty Organization. (2016). *AJP-2.1, edition B, version 1: Allied joint doctrine for intelligence procedures*. Brussels, Belgium: NATO Standardization Office.
O'Brien, B. J. (1989). Words or numbers? The evaluation of probability expressions in general practice. *Journal of the Royal College of General Practitioners, 39*, 98–100.
Okan, Y., Janssen, E., Galesic, M., & Waters, E. A. (2019). Using the Short Graph Literacy Scale to predict precursors of health behavior change. *Medical Decision Making: An International Journal of the Society for Medical Decision Making, 39*(3), 183–195. https://doi.org/10.1177/0272989X19829728
Peterson, D. K., & Pitz, G. F. (1988). Confidence, uncertainty, and the use of information. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*, 85–92.
Redelmeier, D., Koehler, D. J., Liberman, V., & Tversky, A. (1995). Probability judgment in medicine: Discounting unspecified possibilities. *Medical Decision Making, 15*, 227–230.
Reyna, V. F., Nelson, W. L., Han, P. K., & Dieckmann, N. F. (2009). How numeracy influences risk comprehension and medical decision making. *Psychological Bulletin, 135*(6), 943–973. https://doi.org/10.1037/a0017327
Riege, A. H., & Teigen, K. H. (2013). Additivity neglect in probability estimates: Effects of numeracy and response format. *Organizational Behavior and Human Decision Processes, 121*, 41–52. https://doi.org/10.1016/j.obhdp.2012.11.004
Sanbonmatsu, D. M., Posavac, S. S., & Stasney, R. (1997). The subjective beliefs underlying probability overestimation. *Journal of Experimental Social Psychology, 33*, 276–295.
Schwarz, N. (1996). *Cognition and communication: Judgmental biases, research methods, and the logic of conversation*. Hillsdale, NJ: Erlbaum.
Sirota, M., & Juanchich, M. (2015). A direct and comprehensive test of two postulates of politeness theory applied to uncertainty communication. *Judgment and Decision Making, 10*, 232–240.
Teigen, K. H. (1983). Studies in subjective probability III: The unimportance of alternatives. *Scandinavian Journal of Psychology, 24*, 97–105.
Teigen, K. H. (1988). When are low-probability events judged to be "probable"? Effects of outcome-set characteristics on verbal probability judgments. *Acta Psychologica, 67*, 157–174.
Teigen, K. H. (2001). When equal chances = good chances. Verbal probabilities and the equiprobability effect. *Organizational Behavior and Human Decision Processes, 85*, 77–108.
Teigen, K. H. (2022). Judgments by representativeness. In R. F. Pohl (Ed.), *Cognitive illusions: Intriguing phenomena in thinking, judgment, and memory* (3rd. ed., pp. 191–208). London and New York: Psychology Press.
Teigen, K. H., Andersen, B., Alnes, S. L., & Hesselberg, J.-O. (2020). Entirely possible overruns: How people think and talk about probabilistic cost estimates. *International Journal of Managing Projects in Business, 13*(2), 293–311. https://doi.org/10.1108/IJMPB-06-2018-0114
Teigen, K. H., Filkuková, P., & Hohle, S. M. (2018). It can become 5°C warmer: The extremity effect in climate change forecasts. *Journal of Experimental Psychology: Applied, 24*, 3–17. https://doi.org/10.1037/xap0000149
Teigen, K. H., & Jørgensen, M. (2005). When 90% confidence intervals are only 50% certain: On the credibility of credible intervals. *Applied Cognitive Psychology, 19*, 455–475.
Teigen, K. H., Juanchich, M., & Filkuková, P. (2014). Verbal probabilities: An alternative approach. *The Quarterly Journal of Experimental Psychology, 67*(1), 124–146. https://doi.org/10.1080/17470218.2013.793731
Teigen, K. H., Juanchich, M., & Riege, A. H. (2013). Improbable outcomes: Infrequent or extraordinary? *Cognition, 127*(1), 119–139. https://doi.org/10.1016/j.cognition.2012. 12.005
Theil, M. (2002). The role of translations of verbal into numerical probability expressions in risk management: A meta-analysis. *Journal of Risk Research, 5*(2), 177–186. https://doi.org/10.1080/13669870110038179
Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211*(4481), 453–458. https://doi.org/10.1126/science.7455683
Tversky, A., & Kahneman, D. (1982). Judgments of and by representativeness. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 84–98). Cambridge: Cambridge University Press.
Ülkümen, G., Fox, C. R., & Malle, B. F. (2016). Two dimensions of subjective uncertainty: Clues from natural language. *Journal of Experimental Psychology: General, 145*, 1280–1297.
Van der Bles, A. M., van der Linden, S., Freeman, A. L. J., Mitchell, J., Galvao, A. B., Zaval, L., et al. (2019). Communicating uncertainty about facts, numbers and science. *Royal Society open science, 6*, Article 181870. https://doi.org/10.1098/rsos.181870
Van Wallendael, L. R., & Hastie, R. (1990). Tracing the footsteps of Sherlock Holmes: Cognitive representations of hypothesis testing. *Memory and Cognition, 18*, 240–250.
Volden, G. H., & Andersen, B. (2018). The hierarchy of public project governance frameworks: An empirical study of principles and practices in Norwegian ministries and agencies. *International Journal of Managing Projects in Business, 11*(1), 174–197. https://doi.org/10.1108/IJMPB-04-2017-0040
Volden, G. H., & Samset, K. (2017). Governance of major public investment projects: Principles and practices in six countries. *Project Management Journal, 48*(3), 90–109.
Wallsten, T. S., Budescu, D. V., Zwick, R., & Kemp, S. M. (1993). Preferences and reasons for communicating probabilistic information in verbal or numerical terms. *Bulletin of the Psychonomic Society, 31*, 135–138. https://doi.org/10.3758/BF03334162
Weber, E. U., & Hilton, D. J. (1990). Contextual effects in the interpretations of probability words: Perceived base rate and severity events. *Journal of Experimental Psychology: Human Perception and Performance, 16*, 781–789.
Windschitl, P. D., & Wells, G. L. (1996). Measuring psychological uncertainty: Verbal versus numeric methods. *Journal of Experimental Psychology: Applied, 2*, 343–364.

Windschitl, P. D., & Wells, G. L. (1998). The alternative-outcomes effect. *Journal of Personality and Social Psychology, 75*(6), 1411–1423.

Wintle, B. C., Fraser, H., Wills, B. C., Nicholson, A. E., & Fidler, F. (2019). Verbal probabilities: Very likely to be somewhat more confusing than numbers. *PloS one, 14*(4), Article e0213522. https://doi.org/10.1371/journal.pone.0213522

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. https://doi.org/10.1037//0096-3445.124.4.424