Understanding 'Unlikely (20% Likelihood)' Or '20% Likelihood (Unlikely)' Outcomes: The Robustness of the 'Extremity Effect'.

Sarah C. Jenkins[a], Adam J. L. Harris[a] and R. M. Lark[b]

[a] Department of Experimental Psychology, University College London. 26 Bedford Way, London, WC1H 0AP.

[b] British Geological Survey (BGS), Environmental Science Centre, Keyworth, Nottingham NG12 5GG.

Sarah C. Jenkins was funded by an IMPACT studentship from UCL, half funded by the British Geological Survey (BGS). R.M. Lark's contribution is published with permission of the Director, BGS. We thank Dr Charlotte Vye-Brown (BGS) for assistance with creating the vignettes used in Experiments 1 and 3.

A preliminary report of Experiment 1a is found in the Proceedings of the 38th Annual Conference of the Cognitive Science Society (2016).

Correspondence concerning this article should be addressed to Sarah C. Jenkins, Department of Experimental Psychology, University College London, 26 Bedford Way, London, WC1H 0AP, UK. E-mail: s.jenkins.12@ucl.ac.uk

Word count (main text): 8989

Abstract

Calls to communicate uncertainty using mixed, verbal-numerical formats (‘unlikely’ [0-33%]) have stemmed from research comparing mixed with solely verbal communications. Research using the new ‘which outcome’ approach to investigate understanding of verbal probability expressions suggests, however, that mixed formats might convey disadvantages compared to purely numerical communications. When asked to indicate an outcome that is ‘unlikely’, participants often indicate outcomes with a value *exceeding* the maximum value shown, equivalent to a 0% probability (Teigen, Juanchich, & Riege, 2013) – an ‘extremity effect’. Recognising the potential consequences of communication recipients expecting an ‘unlikely’ event to never occur, we extend the ‘which outcome’ work across four experiments, using verbal, numerical and verbal-numerical communication formats, as well as a previously unconsidered numerical-verbal format. We examine how robust the effect is in the context of consequential outcomes and over non-normal distributions. We also investigate whether participants are aware of the inconsistency in their responses from a traditional ‘how likely’ and ‘which outcome’ task. We replicate and extend previous findings, with preference for extreme outcomes (including above maximum values) observed in both verbal and verbal-numerical formats. Our results suggest caution in blanket usage of recently recommended verbal-numerical formats for the communication of uncertainty.

Introduction

Effectively communicating information about risk and uncertainty remains an ongoing challenge for the scientific community. The process relies on recipients of risk communications both understanding the information, and also placing enough trust in it that it will be used in subsequent decision making. Most people do not have in-depth knowledge about, nor experience of, hazards and new technologies (Siegrist, Gutscher, & Earle, 2005). Individuals are therefore reliant on mediated information, which tends to be from an expert source (Sjöberg, 2000). Ensuring the audience understands the information as intended is a universal concern for scientific communications. Scientific forecasts are, however, typically probabilistic (at best). It is not possible to predict with certainty whether a destructive earthquake will occur in a certain place within the next month, for example. A prediction that such an event is ‘unlikely’ does not imply that the event will not occur. Given that an estimate of ‘unlikely’ might be used to describe the probability of events with a 20% likelihood of occurrence (e.g., Theil, 2002), approximately 20% of the time such events will occur. As the prosecution of six experts following the L’Aquila Earthquake in 2009 attests (Cartlidge, 2012), such a lack of certainty is not always well received by the public, resulting in the potential for reduced trust in (and sometimes criminal proceedings against) the scientists who make such predictions. The present paper contributes to our understanding of how probability is communicated and understood by the public by examining different methods of communication, specifically comparing four communication formats: verbal probability expressions (VPEs) (e.g. ‘unlikely’); numerical expressions (e.g. ‘20%’) and mixed expressions in two orders (e.g. ‘unlikely [20% likelihood]’, ‘20% likelihood [unlikely]’)[1].

[1] In line with standard, dictionary definitions, we use ‘likelihood’ as a synonym for ‘probability’ in the present paper, though note that, mathematically, each has a unique and specific definition.

Communication Formats

Budescu and Wallsten (1995) proposed that the choice of format for communicating likelihood information should be governed by the congruence principle: the precision of the communication should be consistent with the degree of certainty that can reasonably be expected for estimates about the event described. In the domain of geological hazards, estimations of events, such as the probability of a large earthquake, might not be precisely quantifiable. In such instances, a specific numerical expression of the probability of this event might be perceived as overly precise. Using a VPE would seem to better represent the uncertainty and underlying imprecision associated with the probability estimate. VPEs are also thought to be easier to understand and more natural for individuals to produce (Budescu & Wallsten, 1987; Erev & Cohen, 1990). There is, however, considerable variability in people's usage and interpretations of VPEs (e.g., Budescu & Wallsten, 1985). In addition to 'natural' inter-individual variability (Beyth-Marom, 1982), interpretations of VPEs are susceptible to contextual and cultural influences (Bonnefon & Villejoubert, 2006; Fischer & Jungermann, 1996; Harris & Corner, 2011; Harris, Corner, Xu, & Du, 2013; Juanchich, Sirota, & Butler, 2012; Teigen & Brun, 1999, 2003; Wallsten, Fillenbaum, & Cox, 1986; Weber & Hilton, 1990). Despite this variability, VPEs continue to be commonly used in a wide range of domains, including accounting (Deloitte, 2008), forensic science (Association of Forensic Science Providers, 2009), and communicating the science of climate change (Mastrandrea et al., 2010).

Studies investigating interpretations of VPEs have typically used the 'how likely' translation approach, whereby people are asked to translate VPEs to corresponding numerical probabilities. However, more recently, Teigen and colleagues have demonstrated that a 'which outcome' (W-O) approach to understanding people's interpretations of VPEs paints rather a different picture (Juanchich, Teigen, & Gourdon, 2013; Løhre & Teigen, 2014; Teigen, Juanchich, & Filkuková, 2014; Teigen et al., 2013). In this approach, participants are shown a histogram depicting a distribution of outcomes and asked to complete probability statements (e.g. “It is unlikely that a battery will last __ hours”) with a value they consider appropriate (see Figure 1). This approach highlights the potential for a large qualitative disparity between the probability statement’s meaning intended by the risk communicator and that which is expected / understood by the recipient of the information. Specifically, Teigen et al. (2013) found that when the term ‘unlikely’ was used to describe outcomes which could be ordered on a unipolar dimension (e.g., battery life), participants interpreted the term as referring to outcomes from the high end of the distribution. Most often participants completed the sentence with a lifetime that *exceeded* the maximum time any sampled battery had lasted – hereafter the ‘extremity effect’. This was despite a mean translation of ‘unlikely’ as 40% in a pre-test.

We believe that findings from the W-O methodology are of importance for our understanding of how people understand risk communications, especially given many real world hazards concern continuous outcomes, such as the extent of coastal erosion or lava flows. Ultimately, the number which a participant assigns to a VPE (e.g., in a translation task) is not of great import. The critical element is how people conceptualise quantitative expressions of risk and, ultimately, use them to guide decision making. Whilst the mechanisms behind the results from alternative methodologies (such as the W-O task) have yet to be identified, our current focus is on extending the communication formats used within this paradigm, given the W-O task can provide additional information over and above typical translation tasks. This information should be heeded for a full understanding of people’s conceptualisations of VPEs, which is especially relevant when making recommendations for effective communication formats. It is therefore worrying that the ‘extremity effect’ is considerably out of line with the prescribed usage of the term ‘unlikely’ in communications of uncertainty. For instance, the

UK’s Defence Intelligence developed a six category translation table in which ‘unlikely’ was translated as 15 – 20% (as cited in Ho, Budescu, Dhami, & Mandel, 2016).

On the one hand, the *use* of ‘unlikely’ to communicate outcomes with a 0% likelihood of occurrence could be argued to be exaggerating the risk, and an appropriate strategy to minimise losses (e.g., by encouraging preventative action) where those associated with an underestimate of the risk are greater than an overestimate (e.g. Batchelor & Peel, 1998; Clatworthy, Peel, & Pope, 2012; Granger, 1969; Harris, Corner, & Hahn, 2009; Lawrence & O’Connor, 2005; Weber, 1994). We propose, however, that the W-O task is informative with respect to how VPEs will be *understood*, and hence acted upon, by recipients of a risk communication. From this perspective, it is easy to foresee how the ‘extremity effect’ could prove deleterious to effective risk communication. If phrases such as ‘unlikely’ are seen as most appropriate for communicating outcomes with no chance of occurring, the mismatch between this and an intended communication of ‘20% likelihood’ could adversely affect confidence in subsequent communications (Jenkins, Harris, & Lark, 2017). More immediately, Teigen and colleagues’ findings suggest there could be extreme consequences for citizens who, upon hearing that the chance of a volcanic lava flow extending as far as their village is ‘unlikely’, may potentially discount the information, believing it will not happen, and thus choose not to evacuate their homes.

The possibility of catastrophic consequences, however, relies on the assumption that the same result from the W-O methodology will be obtained even when one potential outcome is of particular consequence and thus salient. This is of particular relevance given that consequential communications about hazards will, by definition, refer to consequential outcomes. Previous research using the ‘how likely’ methodology suggests that such an assumption might not necessarily hold, as people’s interpretations of VPEs are higher when those VPEs describe a severe outcome than a neutral outcome (e.g., Harris & Corner, 2011;

Weber & Hilton, 1990). More generally, making one outcome particularly consequential in the W-O methodology will enhance its salience, a characteristic present in all real world risk communications[2]. It is possible that increased saliency of a location (or even multiple locations) could lead one to assume all communications are relevant to that particular location. For example, when considering the potential extent of a volcanic lava flow (e.g., Figure 1), the location of a school a certain distance from the volcano might consume the attention of a communicator, such that all communications are deemed to be relevant to the school, in which case the effects reported in Teigen, Juanchich and Filkuková (2014) (see also, Juanchich et al., 2013; Løhre & Teigen, 2014; Teigen & Filkuková, 2013; Teigen, Juanchich, & Riege, 2013) may not occur.

Improving Risk Communications

One commonly proposed solution to reduce mis-communication (from researchers employing the 'how likely' translation methodology) is the use of a mixed format approach to express uncertainty, for example, 'It is unlikely (less than 33%)' (Budescu, Broomell, & Por, 2009; Budescu, Por, & Broomell, 2012; Budescu, Por, Broomell, & Smithson, 2014; Patt & Dessai, 2005; Witteman & Renooij, 2003; see also Harris, Por, & Broomell, 2017). Budescu and colleagues have demonstrated that such a 'verbal-numerical' format increased the correspondence between people's interpretations and the IPCC guidelines, an effect that was replicated in 24 countries (Budescu et al., 2014). A question which arises from the W-O methodology and has, of yet, received little attention (but see Juanchich & Sirota, 2017) is whether the addition of a VPE could actually *harm* the effectiveness of risk communications, in comparison with communications that only use numbers. Ascertaining if the 'extremity effect' is observed with mixed communication formats and consequential scenarios is

[2] Note that because our outcomes are now impactful, our use of the term risk also maintains consistency with its usage in related scientific disciplines (specifically, Earth Sciences, see Rosenbaum & Culshaw, 2003)

imperative given the potential negative consequences for effective communication of risk and uncertainty.

The low probability domain was of particular interest for study because of its prevalence in describing highly consequential outcomes (which are usually unlikely), such as geological hazards. Additionally, negatively directional expressions such as 'unlikely' have been found to induce a large range of interpretations (Smithson, Budescu, Broomell, & Por, 2012).

Overview of Experiments

The present paper therefore aims to further our understanding of the ramifications of the W-O work of Teigen and colleagues (Juanchich et al., 2013; Løhre & Teigen, 2014; Teigen et al., 2014, 2013) by incorporating additional communication formats: mixed expressions (verbal-numerical [V-N] and the previously unstudiednumerical- verbal [N-V] format). We examine how robust the 'extremity effect' is to consequential outcomes (Experiments 1a and 1b), over distributions other than the commonly-used normal (bell-shaped) distribution (Experiment 2) and following another (different) probability estimation task (a translation task, Experiment 3). We also explore the potential influence of numeracy. Examining the effects of using different communication formats is instructive for designing effective future instruments for the communication of risk and uncertainty.

Experiment 1a

The 'extremity effect' has thus far been tested in non-consequential domains such as battery life and mailing letters. We sought to use scenarios featuring outcomes differing in consequence, in order to examine whether the 'extremity effect' would still occur even when potential outcomes were of particular consequence and thus salient. We manipulated the location of the salient site(s) in Experiment 1a.

Method

Participants

One hundred and fifty five participants were recruited for this online experiment via Prolific Academic (www.prolific.ac). They were paid £0.85 upon completion of the experiment. Eight participants were excluded (six due to duplicate IP addresses and two due to lack of consent) leaving a final sample of 147 (83 male) participants, aged 18 – 60 years (*Mdn* = 27).

Design

Communication format (verbal – "unlikely"; numerical – "20% chance" and mixed – "unlikely [20% chance]") was manipulated between-participants. 20% was a plausible value for 'unlikely', given the IPCC's likelihood scale (which suggests using 'unlikely' to communicate probabilities of between 0 and 33%). 20% was also the average numerical translation of 'unlikely' in Theil's (2002) meta-analysis. Numerical point estimates were used rather than range estimates in order to maximise differences between conditions. Scenario (volcano; flood; earthquake; landslide) and saliency (no salient site; close site; far site; multiple sites) were manipulated within-participants. Scenario and saliency were randomised using the Latin Square Confounded method (Kirk, 1969), such that each participant only saw each scenario and each saliency once, but the combinations of these differed systematically across participants.

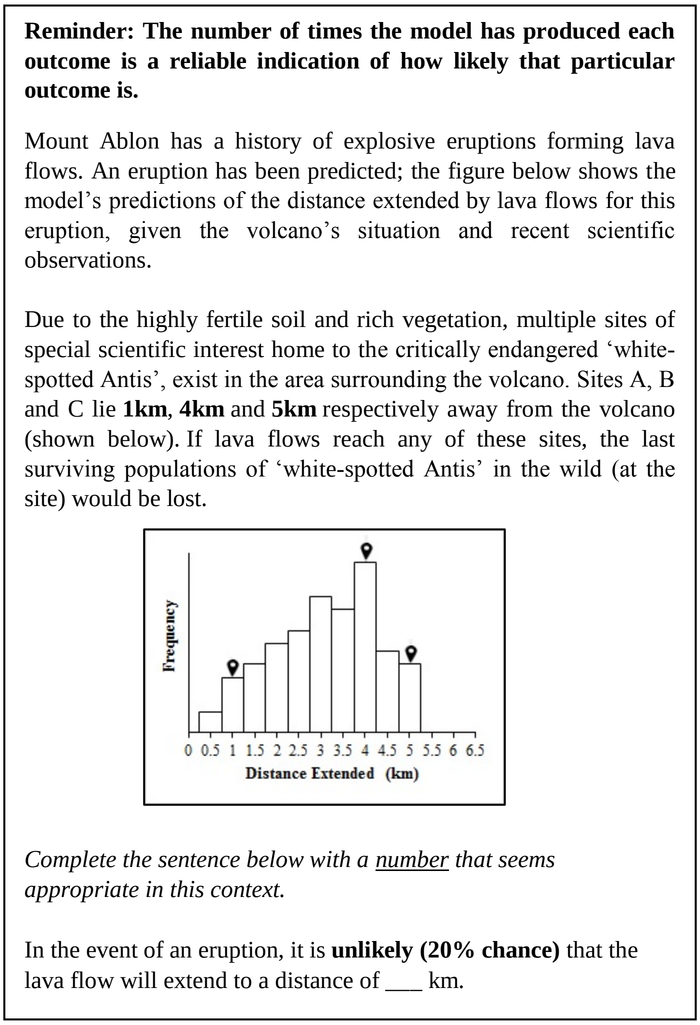
Materials

The introductory text informed participants that they would see reliable projections of a model designed to predict future geological events and be asked to make a series of judgements about these. All participants read four vignettes (developed in conjunction with geologists at the British Geological Survey [BGS] to ensure they reflected plausible real-world situations) describing outcomes of how far lava flows, floodwater, earthquake tremors and debris flows would extend (see Figure 1 for an example and Supplementary Materials 1 for all of the vignettes). Each vignette was illustrated by a histogram which showed the frequency of model-outcomes in each of ten interval bins. The shapes of the distributions were similar and approximately normal across the scenarios, though the distributions in the volcano and floods scenarios had a slightly negative skew. The *y*-axes deliberately featured no values, as we were interested in people's understanding of risk and uncertainty communications, rather than their responses to what they might otherwise have perceived as a mathematical problem. Participants were required to type a numerical response which corresponded to the outcome that was being described.

Saliency was manipulated through the inclusion of sites of particular scientific interest, to which the impact of the geological hazard might extend. These sites were either home to rare plants or critically-endangered animal species (e.g. the last habitat of 'white-spotted Antis' in Figure 1). There were four saliency conditions: no site of interest; one close site (located in the second bin of the histogram); one far site (last bin of the histogram) or multiple sites (second bin, modal bin and last bin, see Figure 1).

Procedure

The experiment was run using Qualtrics (www.qualtrics.com). Participants were first informed about the nature of the experiment and told they could withdraw at any time during the experiment. After consenting to participate, they were asked to indicate their age and gender, before reading the introductory text. The next four screens contained the four vignettes and judgement tasks. Upon completion, participants were given a code to claim their payment, thanked and debriefed.



Reminder: The number of times the model has produced each outcome is a reliable indication of how likely that particular outcome is.

Mount Ablon has a history of explosive eruptions forming lava flows. An eruption has been predicted; the figure below shows the model's predictions of the distance extended by lava flows for this eruption, given the volcano's situation and recent scientific observations.

Due to the highly fertile soil and rich vegetation, multiple sites of special scientific interest home to the critically endangered 'white-spotted Antis', exist in the area surrounding the volcano. Sites A, B and C lie **1km**, **4km** and **5km** respectively away from the volcano (shown below). If lava flows reach any of these sites, the last surviving populations of 'white-spotted Antis' in the wild (at the site) would be lost.

Complete the sentence below with a <u>*number*</u> *that seems appropriate in this context.*

In the event of an eruption, it is **unlikely (20% chance)** that the lava flow will extend to a distance of ___ km.

Figure 1. Example vignette (volcano scenario, multiple salient sites, mixed format).

Results

Effect of Saliency

Outcomes on the *x*-axes were standardised across scenarios by 'binning' responses in all scenarios in relation to the salient points, as if they were in the multiple site condition (1 = below minimum, 2 = minimum, 3 = low saliency point, 4 = between low and mid saliency points, 5 = middle saliency point, 6 = between mid and high saliency points, 7 = high saliency point [maximum] and 8 = above maximum; see Supplementary Materials 2). Given the expected saliency × communication format interaction, an ANOVA was not appropriate because of the Latin Square Confounded Design (Kirk, 1969) and therefore three Kruskal-

Wallis tests were performed. These showed that responses were not significantly affected by saliency in either the verbal, χ^2 (3) = 0.53, *p* = .92, numerical, χ^2 (3) = 4.00, *p* = .26 or mixed format conditions, χ^2 (3) = 3.05, *p* = .39. Whilst we randomised scenario, we also checked for an effect of scenario using three Kruskal-Wallis tests, which showed responses were not affected by scenario in either the verbal, χ^2 (3) = 1.56, *p* = .67, numerical, χ^2 (3) = 5.67, *p* = .13 or mixed format conditions, χ^2 (3) = 1.33, *p* = .72.

Effect of Communication Format

Given the non-significant effect of saliency, we code responses in relation to the bars shown in the histogram (ignoring the salient sites) in all subsequent analyses. For instance, the first bin contains all responses reflecting outcomes to the left of the first histogram bar, the second bin containing all responses reflecting outcomes included in the first histogram bar, and so on (see Figure 1). With ten bars in each histogram and the additional minimum / maximum bins at either end, there were 12 bins in total.

Typical outcomes for ‘unlikely’ were chosen from the higher end of the distribution, from maximum and above maximum observed values (bins 11 & 12) – high amplitude outcomes. In contrast, typical outcomes for ‘20% chance’ tended to correspond to lower values, primarily chosen from the intermediate outcomes[3]. Results for the mixed format fell between those observed with the verbal and numerical formats; outcomes tended to be chosen from the intermediate outcomes, but this did not preclude a sizeable proportion (45.1%) choosing high amplitude outcomes. The contrasting patterns of responses are clearly evidenced in Figure 2.

[3] Effects of communication format were unchanged if responses were binned into five categories (below minimum, minimum, intermediate, maximum and above maximum), as in Teigen et al. (2013).

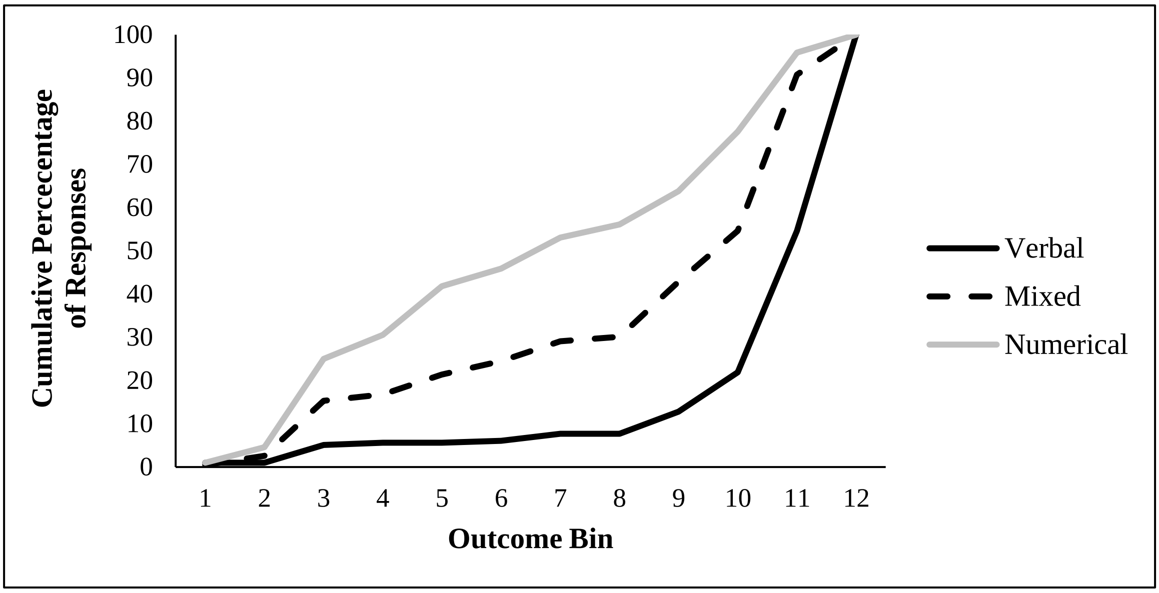


Figure 2. Cumulative Distribution of Responses by Communication Format – Experiment 1a.

To enable comparison with Teigen et al. (2013), responses were coded according to whether they indicated high amplitude outcomes (the maximum value present in the histogram or above – bin 11 or 12) or not. The proportion of responses indicating high amplitude outcomes was highest in the verbal condition, followed by the mixed format condition. The numerical condition had the lowest proportion of responses indicating high amplitude outcomes (see Figure 3)[4].

Further differences between the numerical and mixed formats are observed when one considers that the '20%' in the numerical and mixed communication formats enables the calculation of an objectively correct answer to the statement "there is a 20% chance that the *x* will extend to a distance of __" for the four scenarios. Using the data the histograms were created with, the correct answer was defined cumulatively – representing the forecasted outcome where 20% of forecasted outcomes were that distance or higher. This calculation of the correct answer implies an 'at least' reading of the sentence that is to be completed by

[4] The same format differences are observed even if only 'above maximum' responses are included in analyses. This is also the case for Experiments 1b, 2 and 3.

participants: ‘there is a 20% chance the *x* will extend *at least* ___ km*’*, rather than ‘… *exactly* ___ km’. Whilst some might intuit an ‘exact’ reading of the sentence, searching for an outcome which occurred on *exactly* 20% of occasions, mathematically this is a less appropriate interpretation. The probability of a continuous random variable (such as the distance of a lava flow) taking a single specified value is (strictly) zero. An ‘exact’ reading could be justified were 5 km (for example) given to mean a range from (for example) 4.75 to 5.24 km. We maintain, though, that the ‘at least’ reading is more appropriate, both for the reasons above and because if a lava-flow does extend 5 km from the volcano, sites at any distance up to 5 km along the path of the flow are all affected. The correct answer fell in the 7th bin in three of the four scenarios (for the earthquake scenario it fell in the 8th bin). For these three scenarios, in the numerical condition responses were fairly evenly distributed above and below the correct response (46.9% and 49% respectively), whilst responses in the mixed format condition were more skewed, with 72.8% of responses above the correct response and only 25.2% below[5]. This pattern was also observed in the earthquake scenario. A consideration of responses in the verbal condition was not appropriate here, given the lack of an objective correct response.

[5] The reader might question why a more quantitative comparison of the accuracy of different formats is not included. We see ‘above maximum’ responses as qualitative errors which do not reflect the primary intended meaning of a communication of ‘unlikely’. Given that the correct answer lies towards the upper end of the scale, such errors were not too far from the correct answer. Despite, we argue, being qualitatively inappropriate, these responses (most plentiful in the V-N condition) would be coded as more accurate than some solely quantitative errors in the opposite direction. Further quantitative analysis of accuracy could thus be misleading for purely statistical reasons.

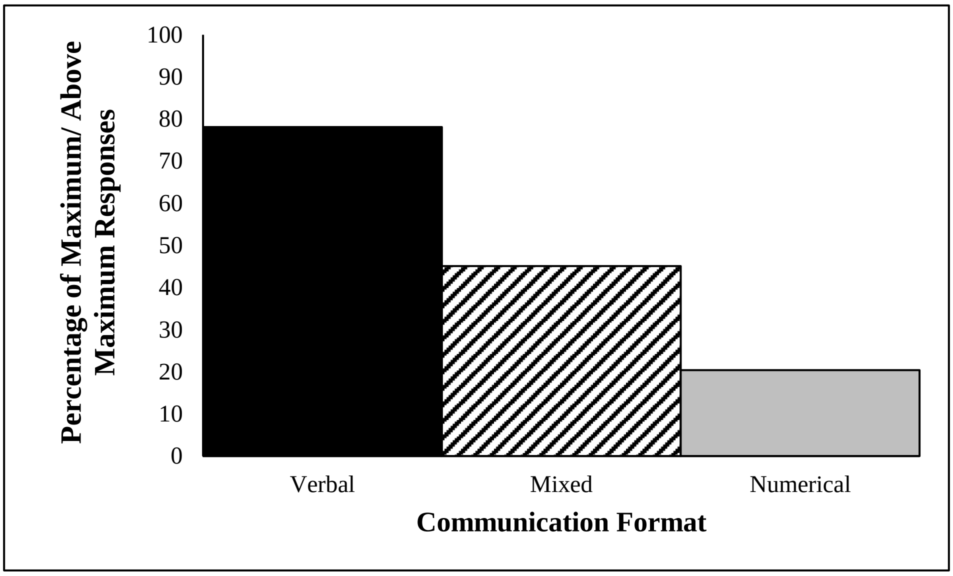


Figure 3. Percentages of Maximum and Above Maximum Responses by Communication Format – Experiment 1a.

Experiment 1b

Although the numerical format was shown to be effective at reducing the proportion of maximum / above maximum responses, it is conceivable that the effectiveness of communication formats could vary across individuals. Less numerate people tend to rely on non-numerical information and have been shown to be vulnerable to the format in which the information is presented (Reyna, Nelson, Han, & Dieckmann, 2010). Furthermore, the influence of numeracy prevails, even in consequential situations (Lipkus, Peters, Kimmick, Liotcheva, & Marcom, 2010). The influence of numeracy will be of particular relevance when considering the effectiveness of using a mixed format, in which the more numerate could be focusing on the numerical expression and the less numerate choosing to focus on the verbal expression. We therefore repeated the experiment in a controlled laboratory session with the addition of a numeracy measure.

Method

Participants

Eighty-three participants were recruited for this online experiment via a first-year undergraduate ‘Introduction to Psychological Experimentation’ class and completed the experiment for course credit. Two participants were excluded due to skipping the consent question, leaving a final sample of 81 (15 male) participants, aged 18 – 20 years (*Mdn* = 19).

Design, Procedure and Materials

As in Experiment 1a. Participants also completed the Berlin Numeracy Test (Cokely, Galesic, Schulz, Ghazal, & Garcia-Retamero, 2012), a series of four questions designed to test numeracy and risk literacy, as the final task in this experiment.

Results

Effect of Saliency

Three Kruskal-Wallis tests were run to investigate if there was an influence of saliency in a) verbal b) numerical and c) mixed format conditions. These showed that responses were not significantly affected by saliency in either the verbal χ^2 (3) = 2.22, p = .53, numerical, χ^2 (3) = 3.53, p = .32 or mixed format conditions χ^2 (3) = 0.55, p = .91. Again, we checked for an effect of scenario. Responses were not significantly affected by scenario in either the verbal χ^2 (3) = 0.42, p = .94, numerical, χ^2 (3) = 1.47, p = .69 or mixed format conditions χ^2 (3) = 0.14, p = .99. In the following analyses, we therefore coded the data as in Experiment 1a.

Effect of Communication Format

The proportion of responses indicating high amplitude outcomes (the maximum / above maximum value present in the histogram) was highest in the verbal condition (60.2%), followed by the mixed format condition (38%). The numerical condition had the lowest proportion of responses (14.7%) indicating high amplitude outcomes. The distribution of responses followed a very similar pattern to those in Experiment 1a (see Figure 4).

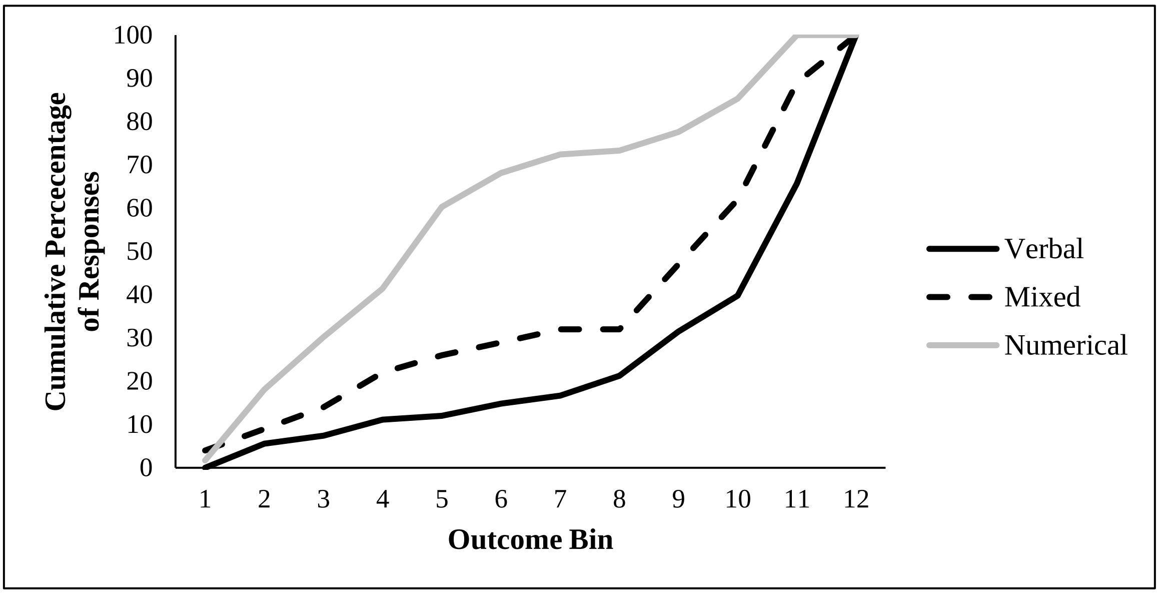


Figure 4. Cumulative Distribution of Responses by Communication Format – Experiment 1b.

As in Experiment 1a, for the scenarios in which the correct answer fell in the 7th bin, responses in the mixed format condition were skewed above the correct answer (69.3%), with 30.7% below. The opposite pattern of responses was found in the numerical condition, in which 28.7% of responses were skewed above the correct answer, with 67.8% below, differing slightly to the even distribution observed in Experiment 1a. A similar pattern of results was found for the earthquake scenario.

Effect of Numeracy

Answers for each question were coded as 1 if correct and 0 if incorrect, such that numeracy scores could range from 0 to 4. The distribution of numeracy scores is shown in Table 1.

Given that we observed no effects of saliency or scenario, we averaged the four responses provided by each participant. A 4 (communication format) × 2 (numeracy – high / low) ANOVA revealed no significant effect of numeracy, *F* (1, 75) = 0.52, *p* = .47, nor a significant interaction between communication format and numeracy, *F* (2, 75) = 0.92, *p* = .40. We finally note that there was a reduction in above maximum responses as numeracy increased, reducing from 42.1% of responses from participants with numeracy scores of 0 or 1, to 15.8% for those with scores of 3 or 4.

Table 1. *Distribution of Numeracy Scores (%) by Experiment. Responses were divided into low and high numeracy (as specified) given the uneven distribution of scores.*

	Numeracy Classification (%)				
	Low		**High**		
Numeracy Score	**0**	**1**	**2**	**3**	**4**
Expt 1a	N/A				
Expt 1b	18.5	28.4	34.6	16.0	2.5
Expt 2	32.9	30.1	18.2	12.1	6.7
Expt 3	N/A				

Discussion

We replicated Teigen et al.'s (2013) results and tested whether these would hold for numerical and mixed format expressions of probability. We found evidence that the tendency to describe outcomes at the very end, or beyond the range of, a distribution generalised to consequential scenarios, wherever the word 'unlikely' was included – the verbal and mixed format conditions. This tendency was not apparent in the numerical condition. We found no effect of including consequential outcomes, that is to say that the presence of a salient site(s) did not affect responses, so this variable is not included in the following experiments.

Whilst the results from Experiment 1b replicate our main findings from Experiment 1a, it is interesting to note the difference in distribution of responses in the numerical condition relative to the correct answer. Whilst responses in the numerical condition in Experiment 1a were evenly split above and below the correct answer, Experiment 1b showed a tendency for more responses being provided below it. A potential to err towards lower estimates might arise because our numerical condition was not purely numerical – '20%' was accompanied by the positive directional term 'chance'. Teigen and Brun (1995, 1999) demonstrated that probability expressions can be distinguished in terms of their directionality. Phrases which have negative directionality (e.g. 'unlikely') focus one's attention on the non-occurrence of the event, whereas those with positive directionality (e.g. 'likely') focus attention on the occurrence of the event. As an expression with positive directionality (Teigen & Brun, 2003), the term 'chance' could thus have led participants in the numerical condition to provide estimates closer to the likely end of the scale (the left). In contrast, those in the mixed format condition may have seen the '20% chance' in parentheses as non-essential information (Walker, 1823), discounted it, and focused on 'unlikely'.

The parenthesis account (Walker, 1823) suggests that the order of the mixed format expression might influence estimates, depending on what information is contained within the parentheses. An order effect would also be in line with Grice's (1975) conversational maxims. The 'maxim of manner' prescribes that utterances should be 'orderly', such that more important information is presented before less important information. Whilst existing research has explored whether a mixed format increases understanding (Budescu et al., 2009, 2014; Witteman & Renooij, 2003), it has not considered the order of the mixed format expression, using only V-N expressions (e.g. 'unlikely [20% likelihood]'), as opposed to N-V expressions (e.g. ‘20% likelihood [unlikely]’). We therefore include this additional communication format in subsequent experiments.

It is also possible that people who focused more on the term ‘unlikely’ in the mixed expression had lower numeracy levels and felt uncomfortable using the ‘20%’ to form their estimates, though we found no evidence to support this assertion in Experiment 1b. Nonetheless, our results provide some suggestion numeracy is having some influence on responses, in that a greater proportion of less numerate participants gave maximum / above maximum responses.

Experiment 2

Not all outcomes for which probabilities must be communicated follow a normal distribution. Examining the ‘extremity effect’ in relation to different distributions can tell us more about the mechanisms behind the effect, for instance whether it is driven by extremeness, or by the frequency of the outcomes. As the name suggests, the effect has been proposed to reflect an extremity preference, with a preference for values from the higher end of the distribution driven by the potential usefulness of being informed about extreme outcomes, irrespective of actual probability (Juanchich et al., 2013). The ‘extremity effect’ was observed for ‘possible’ outcomes over positive skew, negative skew and bi-modal distributions.

Experiment 2 was designed to extend the aforementioned research by examining whether the order of the mixed format expression influenced estimates over varying distributions. It also provided the opportunity to check the generalisability of the results for the numerical and mixed format (V-N) conditions. Given negative skew distributions are uncommon (and less plausible) for geological hazards, we used an altered version of Teigen et al.’s (2013) original battery life scenario. Additionally, we wanted to investigate participants’ understanding of the W-O sentence – are participants completing the sentence with a figure that represents an ‘at least’ or ‘exactly’ interpretation?

Method

Participants

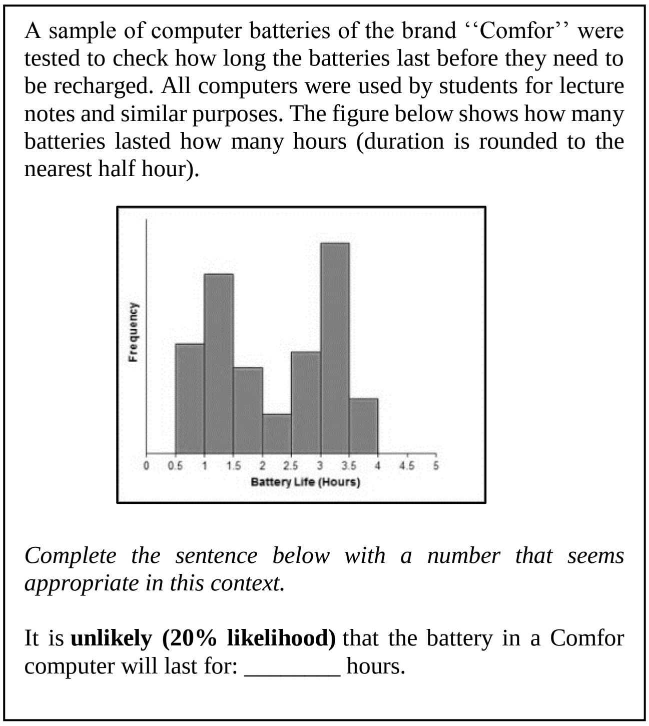
Seven hundred and fifty one participants were recruited for this online experiment via Amazon Mechanical Turk and were paid $0.30 upon completion. 54 cases were removed (due to duplicate IP addresses or for failing the attention check), leaving a final sample of 697 (364 male) participants, aged between 18 – 84 (*Mdn* = 30).

Design

Communication format (verbal – “unlikely”; numerical – “20% likelihood”; V-N – “unlikely [20% likelihood]”; N-V – “20% likelihood [unlikely]”) and distribution (positive skew; negative skew; bi-modal; normal) were manipulated in a 4 × 4 between-participants design. As per Experiment 1, participants were required to type a numerical response which corresponded to the outcome being described.

Materials

The battery life scenario from Teigen et al. (2013, Experiment 3) was used for this experiment. The introductory text informed participants that they would read a short vignette and be asked to make a judgement about it. The text of the vignette was identical to that used originally, though the accompanying histograms illustrating how long the previously tested batteries lasted included seven observed durations of battery life (in comparison to the five in Teigen et al.’s experiment). This faciliated the manipulation of the distribution (see Supplementary Materials 3 for all histograms). A sentence completion task was presented at the bottom of the vignette (see Figure 5 for an example).



A sample of computer batteries of the brand ‘‘Comfor’’ were tested to check how long the batteries last before they need to be recharged. All computers were used by students for lecture notes and similar purposes. The figure below shows how many batteries lasted how many hours (duration is rounded to the nearest half hour).

Frequency
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
Battery Life (Hours)

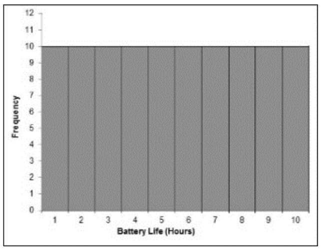
Complete the sentence below with a number that seems appropriate in this context.

It is **unlikely (20% likelihood)** that the battery in a Comfor computer will last for: ________ hours.

Figure 5. Computer Battery Vignette, V-N format, bi-modal distribution – Experiment 2.

After completing the battery task, participants were asked to complete the Berlin Numeracy Test (Cokely et al., 2012). Participants then saw a final histogram, this time of the lifetime of ‘Powerplus’ batteries, in which 10 batteries lasted one hour, 10 batteries lasted two hours, and so on up to 10 hours (see Figure 6). Participants were then asked which of the following statements was ‘most’ correct: “There is a 10% likelihood the battery will last for one hour” versus “There is a 100% likelihood the battery will last for one hour” and why. This task was included in order to establish how participants interpret the W-O sentence. The 10% response denotes the ‘exactly’ interpretation, whilst the 100% response denotes the ‘at least’ interpretation.

A sample of computer batteries of the brand ‘‘Powerplus’’ were tested to check how long they lasted before they need to be recharged. The figure below shows how many batteries lasted how many hours (duration is rounded to the nearest half hour).



Which of these two statements do you think is more correct? Please note there is no correct answer.

- There is a 10% likelihood the battery will last for one hour
- There is a 100% likelihood the battery will last for one hour

Why?

Figure 6. Interpreting the W-O Sentence – ‘correct statement’ question – Experiment 2.

Procedure

The experiment was run using Qualtrics. Participants saw one experimental task screen followed by two screens featuring the Berlin Numeracy Test, followed by the ‘correct statement’ screen. Upon completion, participants were given a code to claim their payment, thanked and debriefed.

Results

Responses were coded into nine bins, with the first bin equivalent to a below minimum response, and the ninth equivalent to an above maximum response. The middle seven represented the seven bars in the histogram (e.g., Figure 5).

The ‘extremity effect’ (with regards to maximum and above maximum responses) was found in all distributions. In each distribution, there was a significant association between communication format and choosing a maximum / above maximum outcome present in the histogram (positive: χ^2 (3) = 31.36, p < .001; normal: χ^2 (3) = 46.36, p < .001; negative: χ^2 (3) = 42.93, p < .001; bi-modal: χ^2 (3) = 57.69, p < .001). The verbal format was consistently associated with the largest proportion of responses indicating high amplitude outcomes and numerical format associated with the smallest proportion (see Figure 7), replicating Experiment 1. We also found a significant association between numeracy and maximum / above maximum responses, χ^2 (4) = 14.15 p < .01, with fewer maximum / above maximum responses as numeracy increased, reducing from 66.8% of responses from participants with numeracy scores of 0 or 1, to 17% for those with scores of 3 or 4.

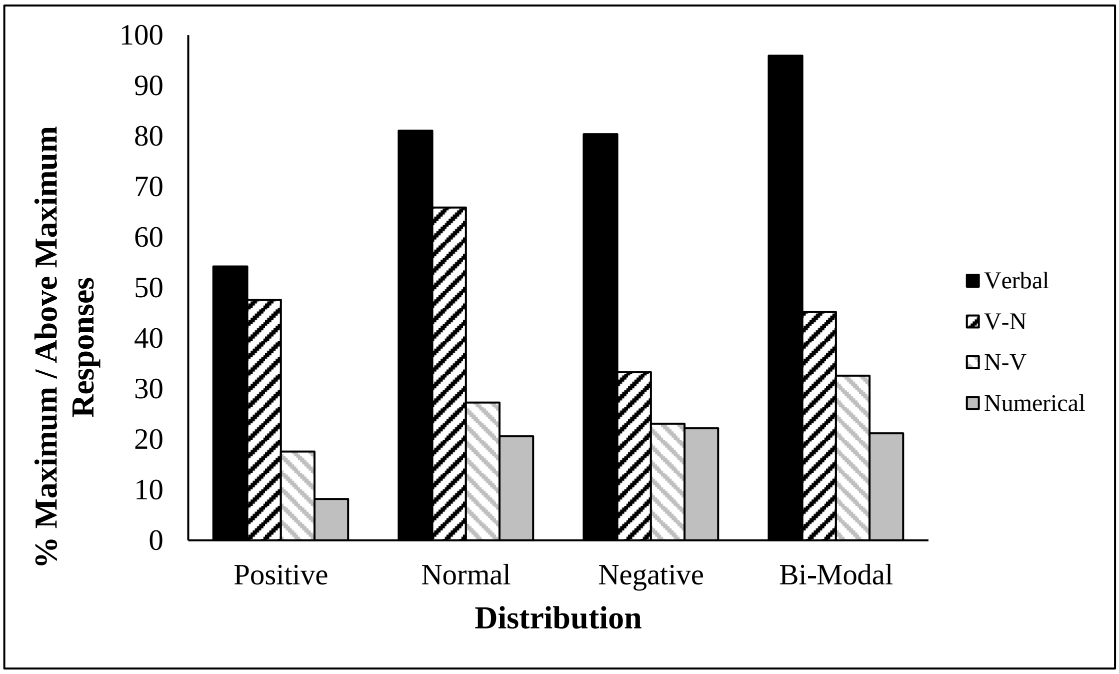


Figure 7. Percentage of Maximum and Above Maximum Responses by Communication Format for Each Distribution – Experiment 2.

Figure 8 shows the mean response by outcome bin for each condition. Consistent with the maximum / above maximum analysis, the highest responses were in the verbal format across all distributions. A 4 (distribution) × 4 (communication format) × 2 (numeracy – high / low) between-subjects ANOVA revealed a significant effect of communication format, *F* (3, 661) = 29.41, $p < .001$, η_p^2= .12, and a marginally significant effect of distribution, *F* (3, 661) = 2.61, $p = .05$, η_p^2= .01, on responses. There was no significant effect of numeracy, *F* (1, 661) = 0.89, $p = .35$. Although the format × distribution interaction approached significance, *F* (9, 681) = 1.85, $p = .06$, η_p^2= .03, the effect of communication format was similar across all distributions (see Figure 8). No other interactions approached significance (all *p*s > .51)[6].

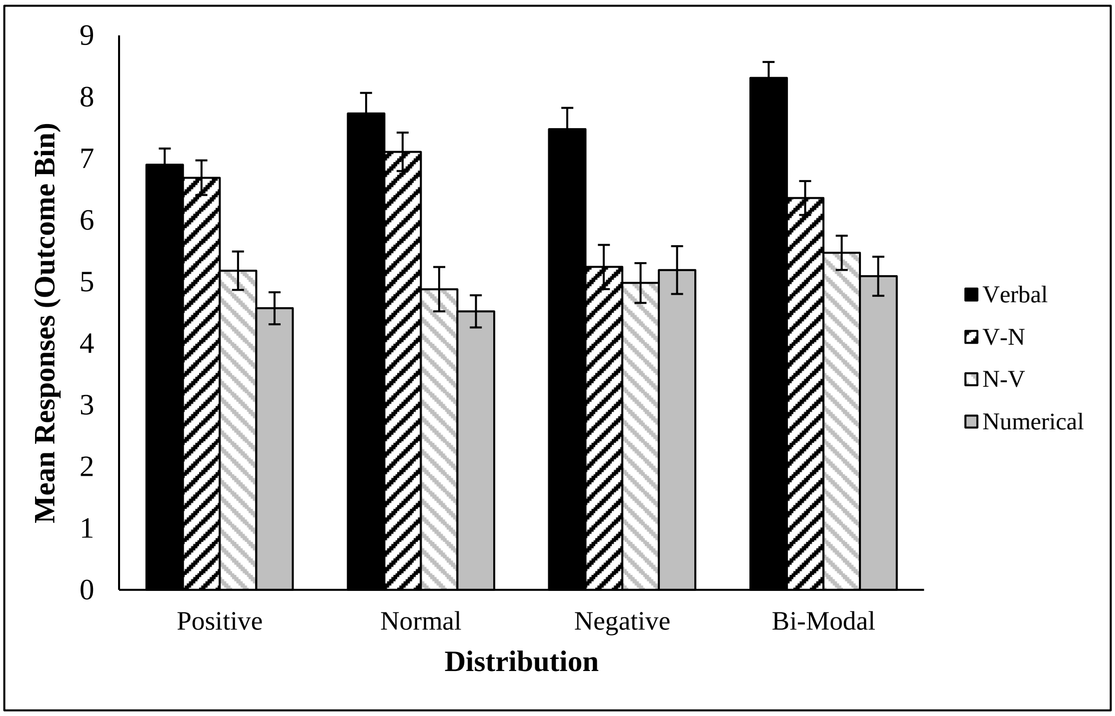


Figure 8. Responses by Communication Format for each Distribution (Error Bars Represent ±1 Standard Error) – Experiment 2.

[6] Cumulative distribution graphs for each distribution can be found in Supplementary Materials 4.

Following Experiments 1a and 1b, we analysed the distribution of responses around the correct answer for the numerical and mixed formats in the normal distribution. The correct answer fell in bin six. As in Experiment 1b, the numerical condition had a higher proportion of responses below the correct answer (61.9%) than above (20.6%). Responses in the V-N condition again replicated, with responses skewed above the correct answer (65.9%) compared to 13.6% below. In the N-V condition 51.5% of responses were below the correct answer and 27.3% were above the correct answer.

Correct Statement

Nearly three-quarters of participants (73.2%) endorsed the statement ‘there is a 100% likelihood the battery will last for one hour’, with 26.8% endorsing the statement ‘there is a 10% likelihood the battery will last for one hour’ as correct, indicating the majority gave the W-O sentence an ‘at least’ interpretation. There was a significant association between numeracy score (high / low) and statement endorsement, χ^2 (1) = 4.65 p < .05, with higher numeracy scores linked to the ‘100% likelihood’ statement.

Discussion

The results of Experiment 2 support the conclusions of Experiment 1 with one additional caveat: the ‘extremity effect’ extends to scenarios wherever the word ‘unlikely’ is included *at the beginning of the probability phrase.* Participants seem to be sensitive to the order of the mixed format expression, attributing greater weight to the information preceding that which is presented in parentheses – in line with Grice’s (1975) stipulation that cooperative utterances will be ‘orderly’. A further experiment (Experiment S1, see Supplementary Materials 5) confirmed that it is the order of the information that drives this effect, rather than the presence of parentheses (e.g., “unlikely – 10-30% likelihood”)[7]. It therefore seems that the

[7] We thank named reviewer, Karl-Halvor Teigen, for suggesting this experiment.

pragmatics of communication are responsible for the ‘extremity effect’, rather than it reflecting a general feature of how people understand probabilities and frequencies. The negative directionality of the expression ‘unlikely’ focuses attention on the non-occurrence of the event and shifts estimates to the ‘unlikely’ (right) end of the scale, whereas ‘20% likelihood’ could be seen as focusing attention on the occurrence of the event and shifting estimates to the ‘likely’ (left) end of the scale.

The majority of participants endorsed the 100% likelihood statement as correct, indicating they interpreted the sentence in the way we have argued is most justifiable. It is interesting to note that more numerate participants provide fewer responses from the high end of the distribution, thus not demonstrating the qualitative inconsistency between indicating an outcome that will *never* happen, and standard uses of ‘unlikely’ to imply probabilities greater than zero.

Overall, Experiments 1 and 2 indicate that the ‘extremity effect’ occurs whenever a VPE is presented first in a probability phrase. The effect seems relatively robust against contextual influences such as the saliency of certain possible outcomes, and the shape of the distribution (similar to Juanchich et al.’s, 2013, findings). Numeracy appears to have only a limited influence on responses.

Experiment 3

The extremity effect observed in W-O tasks (e.g., Experiments 1 & 2) conflicts with responses in translation tasks. Even in a within-participants study, Teigen et al. (2013, Experiment 5b) found participants provided translations of around 30% after having completed an ‘improbable’ sentence with maximum or above maximum responses (0% probability). These results suggest that participants might not be sensitive to the inconsistency between two such responses. We extend this research by providing participants with a second W-O task after the translation task.

We propose that participants who selected an ‘unlikely’ outcome with a 0% frequency of occurrence but yet translated ‘unlikely’ as 20% *should* recognise the inconsistency between their two responses and thus adjust their answers in the second W-O task. We returned to using consequential geological scenarios, having established that our results generalise across different shapes of distribution.

Method

Participants

One hundred and fifty one participants were recuited for this online experiment via Prolific Academic. The were paid £0.50 upon completion of the experiment. One participant was excluded for failing the attention check, leaving a final sample of 150 participants (75 male), aged between 18 – 65 (*Mdn* = 31).

Design

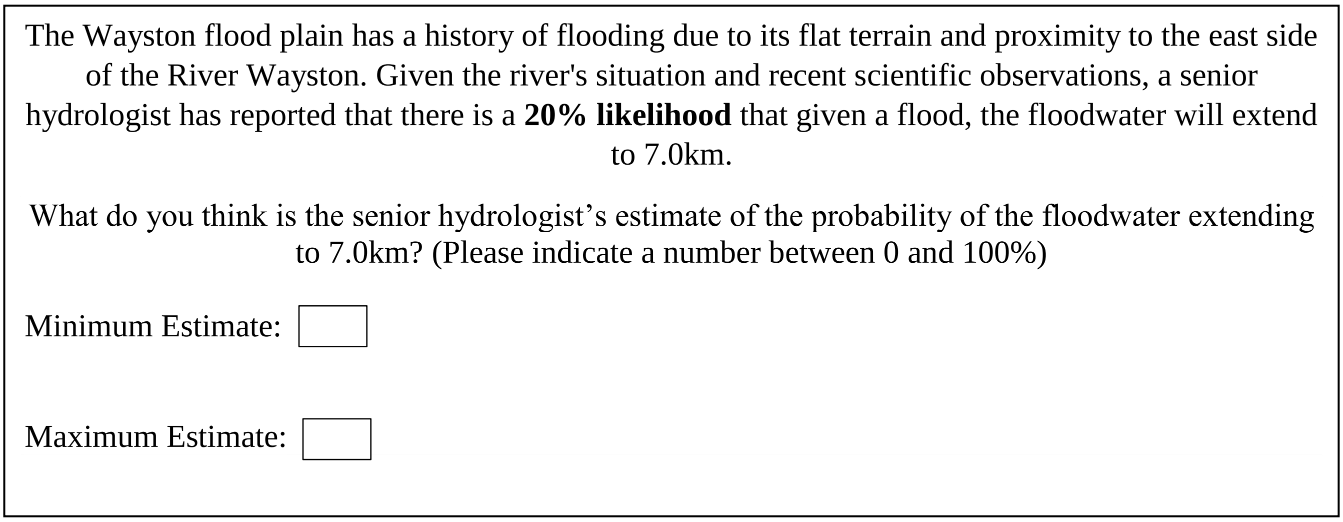
Communication format (verbal – “unlikely”; numerical – “20% likelihood”; V-N – “unlikely [20% likelihood]”; N-V – “20% likelihood [unlikely]”) was manipulated between-participants. Scenario (volcano; flood; earthquake) and W-O task (before and after the translation task) were manipulated within-participants. Scenario was randomised such that each participant only saw each scenario once, but the task they saw it in (e.g. first or second W-O task / translation task) differed systematically across participants. Different scenarios for these tasks were used in order to reduce demand characteristics. No numeracy measure was included in this experiment. For the W-O tasks, participants responded as in Experiments 1 and 2. For the translation task, we asked participants what they thought the probability conveyed by the expert was by indicating a minimum and maximum estimate. This meant the translation made sense for expressions featuring 20% – it would have seemed rather odd to ask for a best estimate of the numerical probability implied by 20% in any of the conditions including the precise number(!)

Materials and Procedure

The experiment was run using Qualtrics. The introductory text informed participants that they would read various predictions about future geological events and be asked to make a series of judgements about these. All participants read three vignettes describing outcomes of how far lava flows, floodwater and earthquake tremors would extend, as in Experiment 1.

Participants completed two W-O tasks. The shape of the distributions were similar and approximately normal across the three (volcano, flood and earthquake) scenarios (see Supplementary Materials 6).

After completing the first W-O task, participants then moved on to a translation task, in which they were required to provide a numerical translation of the probability term used, by giving their minimum and maximum estimates (see Figure 9). They then completed a final W-O task before being thanked and debriefed (see Figure 10 for a flow chart of the procedure).



The Wayston flood plain has a history of flooding due to its flat terrain and proximity to the east side of the River Wayston. Given the river's situation and recent scientific observations, a senior hydrologist has reported that there is a **20% likelihood** that given a flood, the floodwater will extend to 7.0km.

What do you think is the senior hydrologist’s estimate of the probability of the floodwater extending to 7.0km? (Please indicate a number between 0 and 100%)

Minimum Estimate: □

Maximum Estimate: □

Figure 9. Example of translation task (numerical condition).

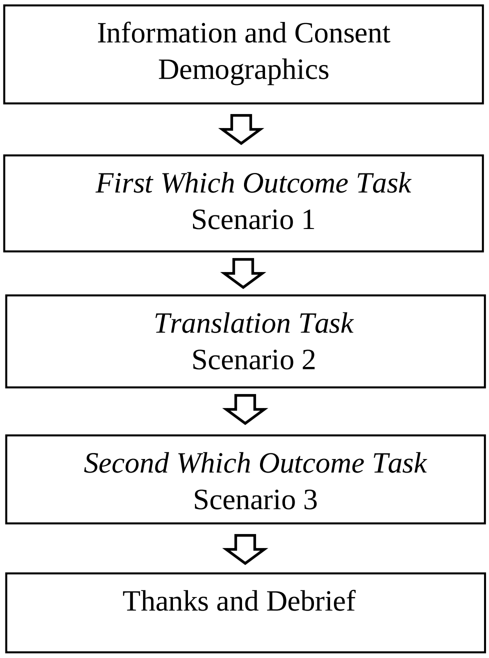


Figure 10. Procedure in Experiment 3. The scenario used as 1, 2 and 3 was randomised across participants.

Results

The first W-O task enabled a replication of the ‘extremity effect’ over all four communication formats with the geological scenarios. The primary aim of the experiment was, however, to examine whether the W-O effect was robust against an interim translation task, which was expected to yield an ‘inconsistent’ response, equating ‘unlikely’ with values greater than 0%. In the analysis, we present the translation task results first, followed by results from the first and then second W-O tasks, before considering the effect of the translation task by examining the difference between the two W-O tasks.

Translation Task

The midpoint of each participant’s minimum and maximum translations was taken as their ‘best estimate’. There was no effect of communication format on estimates, *F* (3, 146) = 0.65, *p* = .59 – estimates in the verbal format (M= 31.38%, SD= 32.83) (Range= 0 – 96.5%, *Mdn*=

15%) were similar to those in the other formats. In the verbal format, only one participant gave an estimate of zero, with 29.7% of participants giving an estimate of 5% or under.

First 'Which Outcome' task

As in Experiment 1, responses on the W-O task were standardised across scenarios by 'binning' responses, in accordance with the order of bars in the histogram. With nine bars in each histogram and an additional minimum / maximum bin at each end, there were 11 bins in total. Responses were similar across scenarios. Similar to previous experiments, the 'extremity effect' replicates, with the proportion of responses indicating high amplitude outcomes largest in the verbal condition, followed by the V-N condition. The numerical condition had the lowest proportion of responses indicating high amplitude outcomes, χ^2 (3) = 58.72, p < .001 (see Figure 11).

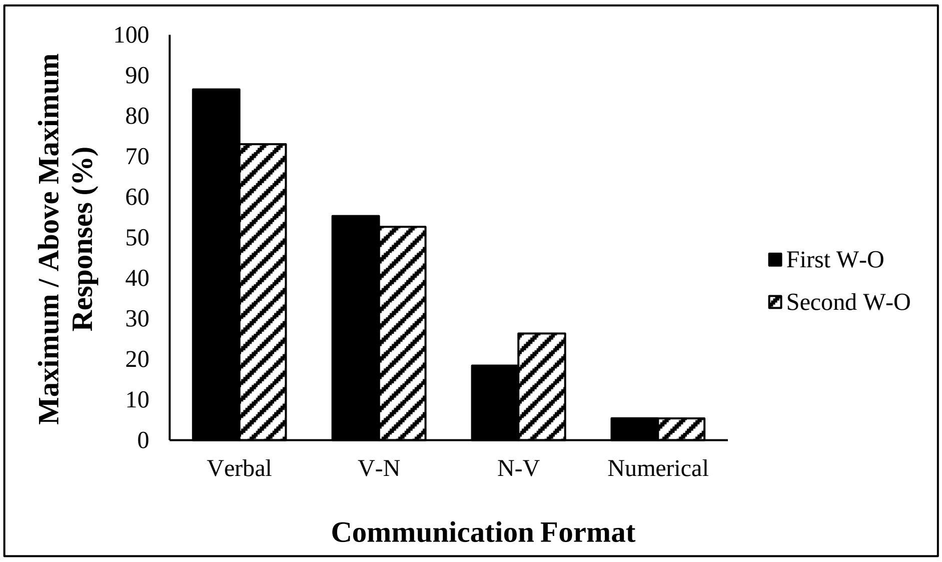


Figure 11. Percentage of Maximum and Above Maximum Responses by Communication Format For First and Second W-O Tasks – Experiment 3.

Figure 12 shows responses by bin, further illustrating the effect of communication format, specifically that the verbal format leads to high amplitude responses. A one-way ANOVA showed a significant effect of communication format *F* (3, 146) = 31.83, *p* < .001,

η_p^2= .40. A REGWQ procedure revealed that responses in the verbal condition (M= 10.00, SE= 0.41) and the V-N condition (M= 9.00, SE= 0.40) were similar. They were significantly higher than responses in the N-V (M= 6.76, SE= 0.40) and numerical (M= 4.89, SE= 0.41) conditions. Responses were significantly different in the numerical and N-V conditions.

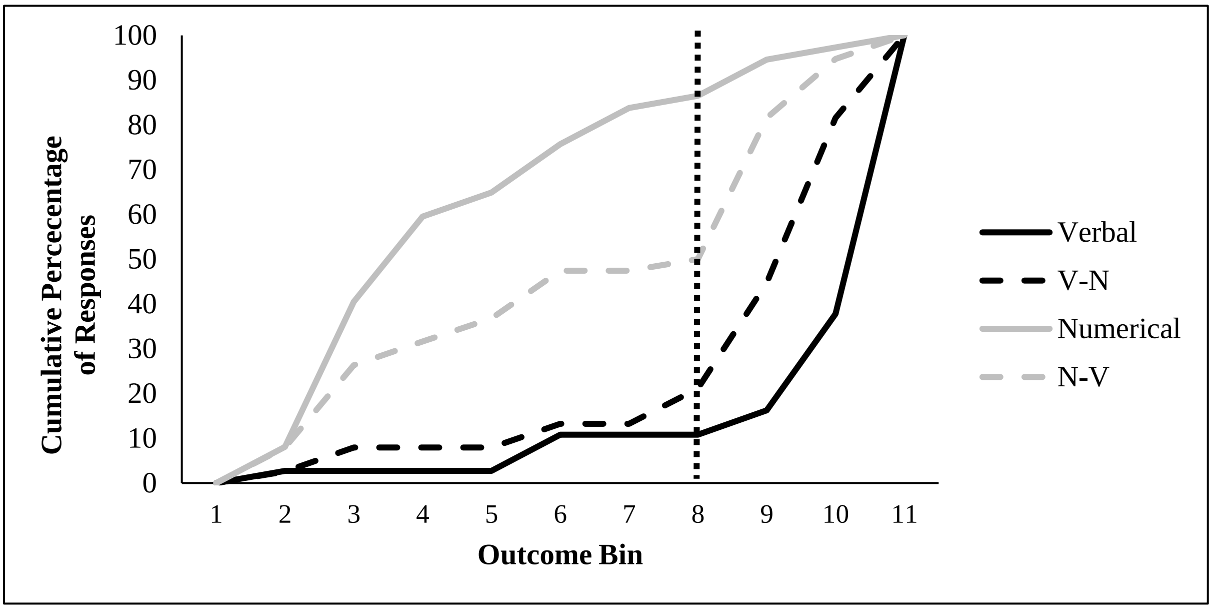


Figure 12. Cumulative Distribution of Responses by Communication Format Pre Translation Task (Dotted Line Represents Outcome Bin Containing Correct Answer) – Experiment 3.

Bin eight included the outcome reflecting a probability of 20% across all three scenarios – the correct answer. As in the previous three experiments, the numerical format had a higher proportion of answers below the correct answer (83.8%) than above (13.5%). Responses from the V-N format again replicated, with a greater proportion of responses above the correct answer (78.9%) compared to below (13.2%). Responses in the N-V format were evenly distributed, with 47.4% below the correct answer and 50% above the correct answer.

Second 'Which Outcome' task

The proportion of responses indicating high amplitude outcomes was again largest in the verbal condition, followed by the V-N condition. The numerical condition had the lowest proportion of responses indicating high amplitude outcomes, χ^2 (3) = 40.91, p < .001, (see Figure 11).

A one-way ANOVA showed a significant effect of communication format, *F* (3, 146) = 15.70, *p* < .001, η_p^2= .24, with the general distribution of responses for each communication format being relatively similar to the first W-O task. A REGWQ procedure revealed that responses in the verbal condition (M= 9.16, SE= 0.47) and the V-N condition (M= 8.29, SE= 0.46) were similar. They were significantly higher than responses in the numerical condition (M= 4.89, SE= 0.47). Responses were not significantly different in the V-N and N-V conditions (M= 7.47, SE= 0.47).

The numerical format had a higher proportion of answers below the correct answer (78.4%) than above (21.6%). Responses from the V-N format again replicated, with a greater proportion of responses above the correct answer (71.1%) compared to below (23.7%). The pattern of responses in the N-V format was less uneven, with 34.2% below the correct answer and 55.3% above the correct answer.

The effect of the translation task

Most notably, a repeated measures ANOVA revealed no significant difference between responses on the first and second W-O tasks *F* (1, 146) = 0.80, *p* = .37. There was an inevitable (given the results above) effect of communication format, *F* (3, 146) = 30.96, *p* < .001, η_p^2= .40. The interaction between W-O task and communication format approached significance, *F* (3, 146) = 2.36, *p* = .07, η_p^2= .05 (see Figure 13).

To investigate directly whether participants continued to endorse maximum / above maximum responses to the same degree in the second W-O task as in the first (given that the

W-O task × communication format interaction term approached significance), four McNemar tests were carried out. No significant differences were observed (all *p*s > .22; see Figure 11).

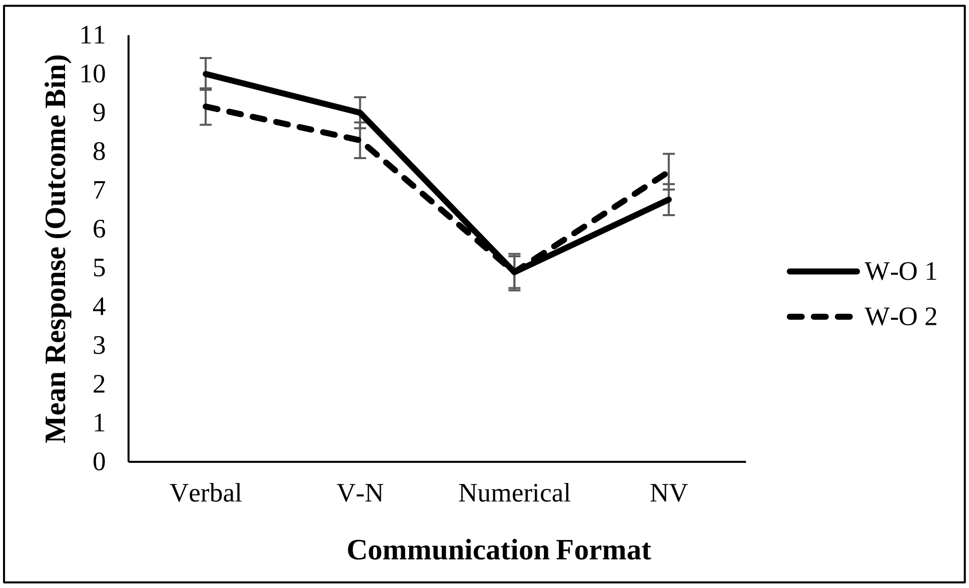


Figure 13. Mean Responses by Communication Format Pre and Post Translation Task (Error Bars Represent ±1 Standard Error) – Experiment 3.

Discussion

The initial W-O task yielded very similar results to previous work: responses in the verbal condition were typically taken from the high end of the distribution. This condition also had the most maximum / above maximum responses. Responses in the V-N condition followed this pattern to a lesser extent. Responses in the numerical condition were typically from the intermediate values in the distribution and had very few maximum / above maximum response (as in the N-V condition). There were no significant differences between responses on the first and second W-O tasks[8].

[8] The reported experiment was a replication of an undergraduate project. The original experiment (thanks to Duyen Tran, Jay See Tow, & Pauline Gordon for data collection) showed a similar pattern of findings but the order effect was not significant. An analysis across the two experiments did, however, reveal a significant order effect, with no moderation by ‘Experiment’.

Results from the translation task showed there were no differences in translations between formats. The mean translation of unlikely was just over 30% – slightly higher than in some previous studies (e.g. Fillenbaum, Wallsten, Cohen, & Cox, 1991; Theil, 2002). Translations were still, however, in line with prescriptions of usage in the real-world (e.g., the Intergovernmental Panel on Climate Change prescribes using 'unlikely' for probabilities of less than 33%).

There was no effect of the translation task on participants' responses in the second W-O task. Participants thus seem unaware of an inconsistency between their responses to 'unlikely' on a translation and W-O task. This might be because participants see the two tasks as unrelated and/or because different processes are involved in the two tasks.

General Discussion

We present four experiments using the W-O methodology to test the robustness of the 'extremity effect' against the presence of a consequential outcome, across differing distributions and communication formats (verbal, numerical and mixed [V-N & N-V]). We also examined whether participants were sensitive to the inconsistencies in their responses in W-O and translation tasks, as well as investigating the potential influence of numeracy.

All experiments yielded consistent results, replicating and extending previous findings, with preference for outcomes at the high end of a distribution (including above maximum values) present in both verbal and V-N conditions. Experiment 3 replicated the effects of communication format, in terms of preferences for the upper end of a distribution in the verbal and V-N conditions, which persisted despite an interim translation task eliciting different responses.

Verbal, Numerical, V-N or N-V Formats?

The differences in responses in the V-N and N-V formats observed in Experiments 2 and 3 (as well as Experiments S1 and S2, see Supplementary Materials), we label an ‘order effect’, in which the first part of the expression is given the most weight. This is in line with Grice's (1975) conversational maxims, particularly the maxim of manner, which states one should be orderly, with the most important information first. The fact that responses in the single format conditions are not exactly the same as their mixed format counterparts implies that the secondary information is underweighted (relative to the initial information) rather than being wholly disregarded. Furthermore, the perceived usefulness (and function) of the secondary information varies between the two mixed format conditions. In the V-N condition, participants may have an initial estimate which is then anchored upwards or downwards with the following numerical expression. In contrast, in the N-V condition, the participant has a very clear number to start with.

That the ‘extremity effect’ continued to occur in the V-N format is of particular relevance to current literature, given recent recommendations to use a mixed format approach to express uncertainty (Budescu et al., 2009, 2012, 2014; Patt & Dessai, 2005; Witteman & Renooij, 2003). The present results suggest that simply including numerical translations alongside VPEs will potentially be of only limited benefit in improving the effectiveness of risk communications. Whilst using a V-N format has been found to increase the differentiation of participants’ interpretations of VPEs, as well as increasing the level of agreement with IPCC guidelines (Budescu et al., 2014), our findings highlight a potential downside of the V-N format, as well as a *potential* upside of the N-V format. Whilst the numerical [V-N] format led to a preponderance of responses below [above] the correct answer, responses in the N-V format were more evenly distributed above and below the correct answer. Of most consequence, however, was that the V-N format showed an increased endorsement of outcomes with a 0%

frequency of occurrence compared to the numerical and N-V formats. This endorsement could be seen as beneficial from the perspective of the communicator, who recognises the costs of underestimating the risk, in line with the asymmetric loss function account (Harris et al., 2009; Weber, 1994), and sees overestimation of a lava flow as a way of motivating action. From the perspective of the recipient, however, such an endorsement is problematic, given that organisations such as the IPCC use 'unlikely' to mean less than 33%, rather than 0%. A failure to act (e.g., evacuate a danger area) based on a 0% interpretation could therefore have life-threatening consequences. Ultimately, the appropriate choice of communication format will depend on the purpose of the communication, and how important minimisation of the 'extremity effect' is to the communicator.

Explaining the 'extremity effect'

Our findings clearly demonstrate that communication format influences interpretations of risk communications. The fact that the 'extremity effect' persists in the mixed V-N format, but not to phrases where the numerical expression comes first indicates that the effect is not related to how people understand probability and frequencies in general, but rather the pragmatics of communication. Although the precise mechanisms underlying the 'extremity effect' remain unknown, we suggest further research should further explore a pragmatic-based account, specifically relating to the directionality of the VPE (Teigen, 1988; Teigen & Brun, 1995, 1999). The term 'unlikely' is negatively directional, focusing one's attention on the non-occurrence of the event. It seems plausible that it is this focus which leads participants to select outcomes from the distribution which have never occured. Further support for such an explanation comes from demonstrations that communications describing a future event as 'unlikely' are perceived as more incorrect following the occurrence of said event than communications describing its likelihood numerically (Jenkins, Harris, & Lark, 2017).

Effect of Numeracy

Compared to more numerate decision makers, those of lower numeracy are “left with information that is less complete and less understood” (Peters et al., 2006, p. 412). Varying numeracy levels (and comfort with numerical information) have therefore been cited as a reason for presenting information in a mixed format (Witteman & Renooij, 2003). We originally conjectured that the high amplitude responses present in 45% of cases in the V-N condition of Experiment 1a (see Figure 3) could have been due to numeracy levels, with responses varying according to two different focuses. The less numerate participants may have felt uncomfortable using ‘20% chance’ to form their estimates and thus focused on ‘unlikely’, but the more highly numerate may have focused on the opposite, which could have explained why the ‘extremity effect’ persisted, but less frequently. However, the contrasting pattern of responses in the N-V conditions from our subsequent experiments suggest that the effect observed in this paper is one of order, where whatever information is presented first is deemed most important.

Nevertheless, over all communication formats and all experiments with a numeracy measure, more numerate participants gave fewer maximum and above maximum responses. This result challenges the notion that VPEs are particularly useful when communicating with less numerate audiences (Gurmankin, Baron, & Armstrong, 2004), instead indicating that they could do most harm for such audiences. We speculate this could be because they encourage a non-mathematical interpretation of a mathematical question, leaving participants susceptible to the effects of the directionality of the expression.

Implications for Trust in Risk Communication

The results from W-O research continue to highlight the variability in people’s understanding of information about risk and certainty, even if the proposed solution of a mixed (V-N) format is used. Our results show the ‘extremity effect’ is pervasive and robust, meaning there could be important (and yet to be researched) consequences for the perceived reliability and credibility of the communicator. For instance, an expert who uses ‘unlikely’ to mean 20% will quickly lose the trust of an audience if that audience expects ‘unlikely’ to refer to outcomes which never happen (Jenkins et al., 2017). This loss of trust could be further compounded by people’s misunderstandings of uncertainty, with uncertainty often perceived by the public as an ‘indicator of ignorance’ (Lewandowsky, Ballard, & Pancost, 2015). Yet scientific forecasts are probabilistic (at best) and thus it is not possible to predict with certainty the probability of a landslide on a given day (for example). Even if such an event is predicted to be ‘unlikely’ to occur, the very fact that the outcome is not certain means that it could still happen. If it does occur, the prediction could be perceived as an ‘erroneous’ prediction’, which will have implications for the perceived credibility of the communicator (Jenkins et al., 2017). Future work should therefore seek to gain a deeper understanding of the effects of using different communication formats, such that the public’s trust in science can be preserved and cultivated (Nature, 2010).

The aforementioned studies used point estimates in the numerical formats. This further extended the work of Budescu and colleagues (e.g., Budescu et al., 2009) who, in their mixed format condition, used more range-like expressions, such as ‘unlikely (< 33%)’. The present results are, however, applicable to real-world risk communications where range estimates are most suitable, as the results reported replicate if numerical ranges are used instead of numerical point estimates (see Experiment S2, Supplementary Materials 7). A potential critique of the work presented thus far is its focus on one low probability VPE, ‘unlikely’. This decision was made in light of the larger range of interpretations given to negative directionality expressions (Smithson et al., 2012) and the relevance of using such phrases to describe highly consequential events (which are usually unlikely) such as geological hazards. However, previous investigations have shown that effects observed with ‘unlikely’ are also observed with positively directional VPEs such as ‘possible’ and ‘chance of occurring’ (Teigen et al., 2014). It therefore seems reasonable to expect similar results to Experiments 1, 2 and 3 for other low probability expressions. We do, however, recognise that it is a more open question as to whether these results will extend to high probability VPEs such as ‘quite probable’ (see Teigen, Juanchich, & Filkuková, 2014).

Conclusion

The present research provides an exploration of the effects of communication format on the understanding of risk and uncertainty communications. We provide further evidence of the potential for discrepancies between the way these expressions are typically used in formal risk communications (e.g., from the UK Defence Intelligence or the IPCC) and the way they are understood by recipients. Whilst it is generally acknowledged that there is no ‘optimal’ presentation format, and no single ‘fix’ for communications of risk and uncertainty (Budescu et al., 2012), identifying instances in which the communication format has a significant impact on audience’s understanding is key to improvement. Our findings show that the ‘extremity effect’ extends to risk communications which use a V-N approach and appears robust over differing distributions as well as the presence of consequential outcomes. These findings, together with observations that the effect is reduced in the N-V format, suggest that refinement of recent recommendations to use a mixed (V-N) communication format is required.

References

Association of Forensic Science Providers. (2009). Standards for the formulation of evaluative forensic science expert opinion. *Science & Justice*, *49*(3), 161–164.

Batchelor, R., & Peel, D. A. (1998). Rationality testing under asymmetric loss. *Economics Letters*, *61*, 49–54.

Beyth-Marom, R. (1982). How Probable is Probable? A Numerical Translation of Verbal Probability Expressions. *Journal of Forecasting*, *1*, 257–269.

Bonnefon, J. F., & Villejoubert, G. (2006). Tactful or doubtful? Expectations of politeness explain the severity bias in the interpretation of probability phrases. *Psychological Science*, *17*(9), 747–751.

Budescu, D. V, Broomell, S. B., & Por, H. H. (2009). Improving communication of uncertainty in the reports of the intergovernmental panel on climate change. *Psychological Science*, *20*(3), 299–308.

Budescu, D. V, Por, H. H., & Broomell, S. B. (2012). Effective communication of uncertainty in the IPCC reports. *Climatic Change*, *113*(2), 181–200.

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. Retrieved from http://www.scopus.com/inward/record.url?eid=2-s2.0-84901609016&partnerID=tZOtx3y1

Budescu, D. V, & Wallsten, T. S. (1985). Consistency in interpretation of probabilistic phrases. *Organizational Behavior and Human Decision Processes*, *36*(3), 391–405.

Budescu, D. V, & Wallsten, T. S. (1987). Subjective estimation of precise and vague uncertainties. In G. Wright & P. Ayton (Eds.), *Judgmental Forecasting* (pp. 63–82).

Oxford, England: John Wiley & Sons, Ltd.

Budescu, D. V, & Wallsten, T. S. (1995). Processing Linguistic Probabilities: General Principles and Empirical Evidence. *Psychology of Learning and Motivation*, *32*(2), 275–318.

Cartlidge, E. (2012). Earthquake Experts Convicted of Manslaughter.

Clatworthy, M. A., Peel, D. A., & Pope, P. F. (2012). Are analysts' loss functions asymmetric? *Journal of Forecasting*, *31*(8), 736–756.

Cokely, E. T., Galesic, M., Schulz, E., Ghazal, S., & Garcia-Retamero, R. (2012). Measuring risk literacy: the Berlin numeracy test. *Judgment and Decision Making, 7*(1), 25–47.

Deloitte. (2008). Assets held for sale and discontinued operations — A guide to IFRS 5.

Erev, I., & Cohen, B. L. (1990). Verbal versus numerical probabilities: Efficiency, biases, and the preference paradox. *Organizational Behavior and Human Decision Processes, 45*(1), 1–18.

Fillenbaum, S., Wallsten, T. S., Cohen, B. L., & Cox, J. A. (1991). Some Effects of Vocabulary and Communication Task on the Understanding and Use of Vague Probability Expressions. *The American Journal of Psychology*, *104*(1), 35–60.

Fischer, K., & Jungermann, H. (1996). Rarely occurring headaches and rarely occurring blindness: Is rarely = rarely? *Journal of Behavioral Decision Making*, *9*, 153–172.

Granger, C. W. J. (1969). Prediction with a Generalized Cost of Error Function. *Journal of the Operational Research Society, 20*(2), 199–207.

Grice, P. H. (1975). Logic and Conversation. In *Syntax and Semantics 3: Speech Acts* (pp. 41–58). San Diego, CA: Academic Press.

Gurmankin, A. D., Baron, J., & Armstrong, K. (2004). The effect of numerical statements of risk on trust and comfort with hypothetical physician risk communication. *Medical Decision Making*, *24*(3), 265–271.

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*(6), 1571–8.

Harris, A. J. L., Corner, A., & Hahn, U. (2009). Estimating the probability of negative events. *Cognition*, *110*(1), 51–64.

Harris, A. J. L., Corner, A., Xu, J., & Du, X. (2013). Lost in translation? Interpretations of the probability phrases used by the Intergovernmental Panel on Climate Change in China and the UK. *Climatic Change*, *121*(2), 415–425.

Harris, A. J. L., Por, H. H., & Broomell, S. B. (2017). Anchoring climate change communications. *Climatic Change*, *140*(3–4), 387–398.

Ho, E. H., Budescu, D. V, Dhami, M. K., & Mandel, D. R. (2016). Improving the communication of uncertainty in climate science and intelligence analysis. *Behavioral Science & Policy*, *1*(2), 43–55.

Jenkins, S. C., Harris, A. J. L., & Lark, R. M. (2017). Maintaining Credibility When Communicating Uncertainty : The Role of Communication Format. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. J. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 582–587). Austin, TX: Cognitive Science Society.

Juanchich, M., & Sirota, M. (2017). How much will the sea level rise ? Outcome selection and subjective probability in climate change predictions. *Journal of Experimental Psychology: Applied*, Advance online publication.

Juanchich, M., Sirota, M., & Butler, C. L. (2012). The perceived functions of linguistic risk quantifiers and their effect on risk, negativity perception and decision making. *Organizational Behavior and Human Decision Processes, 118*(1), 72–81.

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.

Kirk, R. E. (1969). *Experimental Design: Procedures for the Behavioral Sciences.* Belmont: CA: Brooks/Cole.

Lawrence, M., & O'Connor, M. (2005). Judgmental forecasting in the presence of loss functions. *International Journal of Forecasting, 21*(1), 3–14.

Lewandowsky, S., Ballard, T., & Pancost, R. D. (2015). Uncertainty as knowledge. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *373*(2055), 1–11.

Lipkus, I. M., Peters, E. M., Kimmick, G., Liotcheva, V., & Marcom, P. (2010). Breast Cancer Patients' Treatment Expectations after Exposure to the Decision Aid Program Adjuvant Online: The Influence of Numeracy. *Medical Decision Making*, *30*(4), 464–473.

Lipkus, I. M., Samsa, G., & Rimer, B. K. (2001). General performance on a numeracy scale among highly educated samples. *Medical Decision Making : An International Journal of the Society for Medical Decision Making*, *21*(1), 37–44.

Løhre, E., & Teigen, K. H. (2014). How fast can you (possibly) do it, or how long will it (certainly) take? Communicating uncertain estimates of performance time. *Acta Psychologica*, *148*, 63–73.

Mastrandrea, M. D., Field, C. B., Stocker, T. F., Edenhofer, O., Ebi, K. L., Frame, D. J., … Zwiers, F. W. (2010). *Guidance Note for Lead Authors of the IPCC Fifth Assessment*

Report on Consistent Treatment of Uncertainties. Retrieved from http://www.ipcc.ch

Nature. (2010). A Question of Trust. *Nature*, *466*(7302).

Patt, A. G., & Dessai, S. (2005). Communicating uncertainty: Lessons learned and suggestions for climate change assessment. *Comptes Rendus Geoscience*, *337*(4), 425–441.

Peters, E. M., Västfjäll, D., Slovic, P., Mertz, C. K., Mazzocco, K., & Dickert, S. (2006). Numeracy and Decision Making. *Psychological Science, 17*, 407–413.

Reyna, V. F., Nelson, W. L., Han, P. K. J., & Dieckmann, N. (2010). How Numeracy Influences Risk Comprehension and Medical Decision Making. *Psychological Bulletin*, *135*(6), 943–973.

Rosenbaum, M. S., & Culshaw, M. G. (2003). Communicating the risks arising from geohazards. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, *166*(2), 261–270.

Siegrist, M., Gutscher, H., & Earle, T. C. (2005). Perception of risk: the influence of general trust, and general confidence. *Journal of Risk Research*, *8*(2), 145–156.

Sjöberg, L. (2000). Factors in risk perception. *Risk Analysis*, *20*(1), 1–11.

Smithson, M., Budescu, D. V, Broomell, S. B., & Por, H. H. (2012). Never say “not”: Impact of negative wording in probability phrases on imprecise probability judgments. *International Journal of Approximate Reasoning*, *53*(8), 1262–1270.

Teigen, K. H. (1988). The language of uncertainty. *Acta Psychologica*, *68,* 27–38.

Teigen, K. H., & Brun, W. (1995). Yes, but it is uncertain: Direction and communicative intention of verbal probabilistic terms. *Acta Psychologica*, *88*(3), 233–258.

Teigen, K. H., & Brun, W. (1999). The Directionality of Verbal Probability Expressions:

Effects on Decisions, Predictions, and Probabilistic Reasoning. *Organizational Behavior and Human Decision Processes*, *80*(2), 155–190.

Teigen, K. H., & Brun, W. (2003). Verbal Probabilities: A Question of Frame? *Journal of Behavioral Decision Making*, *16*(1), 53–72.

Teigen, K. H., & Filkuková, P. (2013). Can & Will: Predictions of What Can Happen are Extreme, but Believed to be Probable. *Journal of Behavioral Decision Making*, *26*(1), 68–78.

Teigen, K. H., Juanchich, M., & Filkuková, P. (2014). Verbal probabilities: An alternative approach. *The Quarterly Journal of Experimental Psychology*, *67*(1), 124–146.

Teigen, K. H., Juanchich, M., & Riege, A. H. (2013). Improbable outcomes: Infrequent or extraordinary? *Cognition*, *127*(1), 119–139.

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.

Walker, J. (1823). *A Rhetorical Grammar or Course of Lessons in Elocution*. Oxford, England: Oxford University Press.

Wallsten, T. S., Fillenbaum, S., & Cox, J. A. (1986). Base rate effects on the interpretations of probability and frequency expressions. *Journal of Memory and Language*, *25*(5), 571–587.

Weber, E. U. (1994). From Subjective Probabilities to Decision Weights: The Effect of Asymmetric Loss Functions on the Evaluation of Uncertain Outcomes and Events. *Psychological Bulletin*, *115*(2), 228–242.

Weber, E. U., & Hilton, D. J. (1990). Contextual effects in the interpretations of probability words: Perceived base rate and severity of events. *Journal of Experimental Psychology: Human Perception and Performance*, *16*(4), 781–789.

Witteman, C. L. M., & Renooij, S. (2003). Evaluation of a verbal-numerical probability scale. *International Journal of Approximate Reasoning*, *33*(2), 117–131.

Supplementary Materials

Contents

1. Geological Scenarios

Volcano

Mount Ablon has a history of explosive eruptions forming lava flows. An eruption has been predicted; the figure below shows the model's predictions of the distance extended by lava flows for this eruption, given the volcano's situation and recent scientific observations.

Flooding

The Wayston flood plain has a history of flooding due to its flat terrain and proximity to the east side of the River Wayston. A flood has been predicted; the figure below shows the model's predictions of the distance extended by floodwater for this flood, given the river's situation and recent scientific observations.

Earthquake

The uninhabited area of Griffinton lies on an active fault line and has a history of experiencing earthquake activity, resulting in tremors which can be felt in various areas. A large earthquake has been predicted; the figure below shows the model's predictions of the distance extended by the tremors for this earthquake, given the fault line's situation and recent scientific observations.

Landslide

A site on the Copeland Pass has been identified by geologists as susceptible to debris flow, with evidence of past flow of material occurring in certain weather conditions. A landslide has been predicted; the figure below shows the model's predictions of the length extended by the debris flow for this landslide, given its situation and recent scientific observations.

2. Geological Scenario Figure Examples- Experiment 1
(Correct answer in brackets)

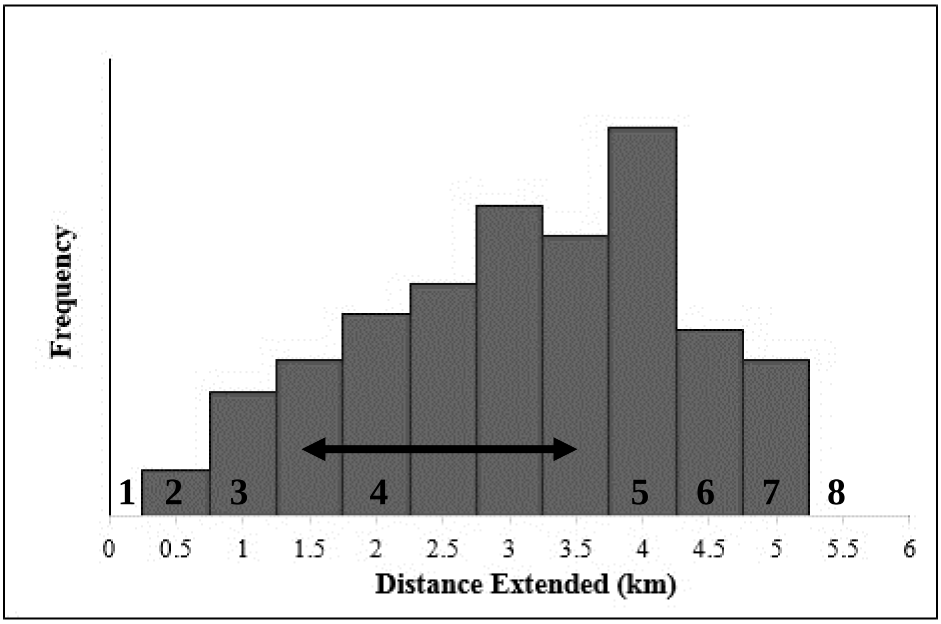


Figure S 1. Volcano – No Salient Site (4 km). Bold numbers represent outcome bin.

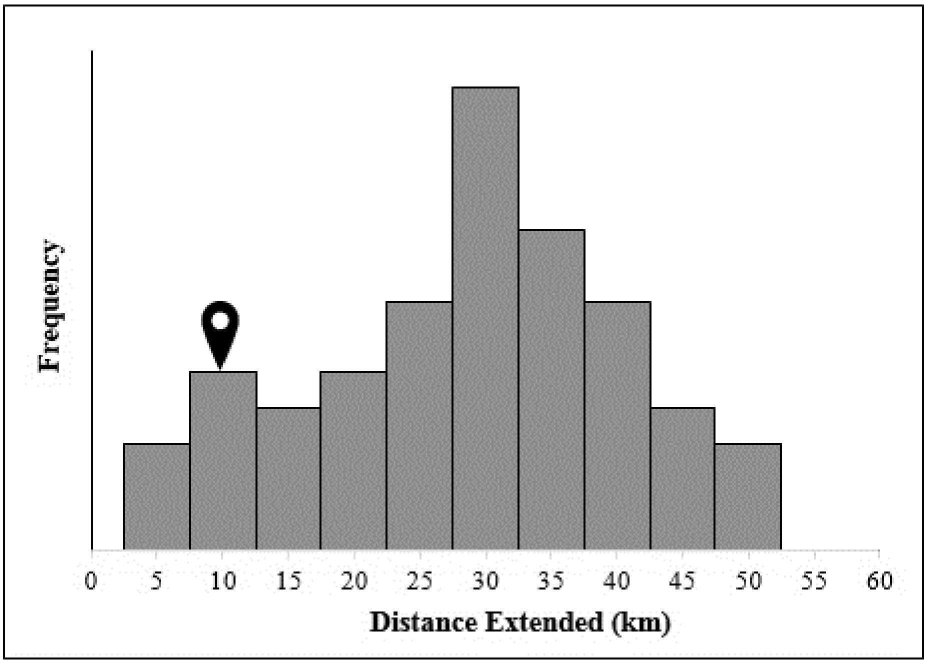


Figure S 2. Earthquake – Close Salient Site (40 km).

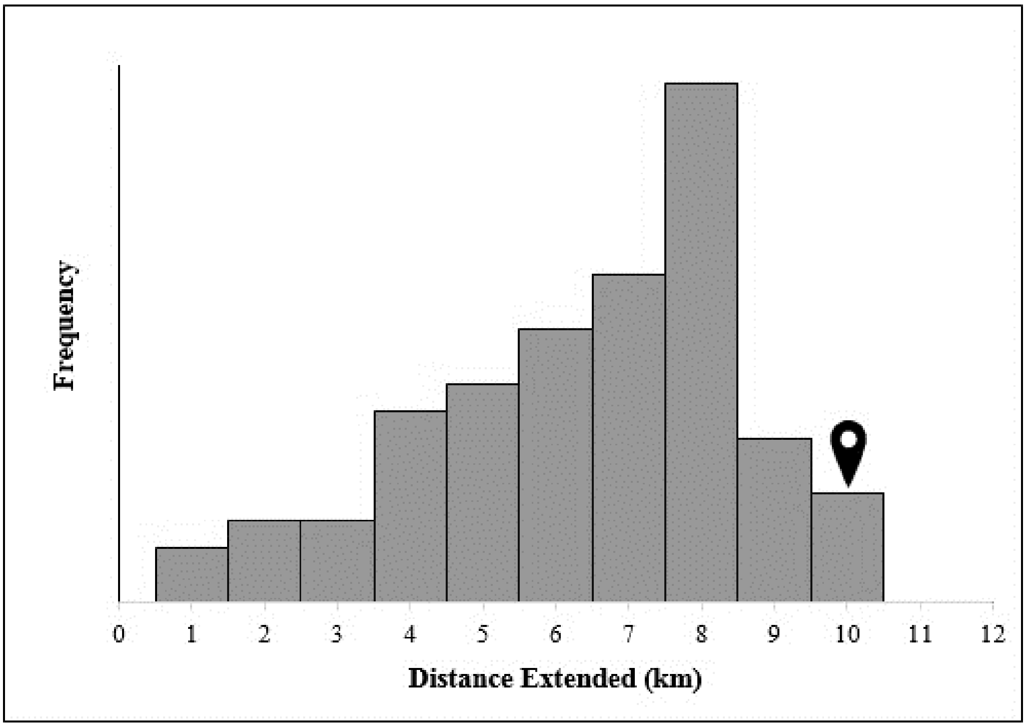


Figure S 3. Flooding – Far Salient Site (8 km).

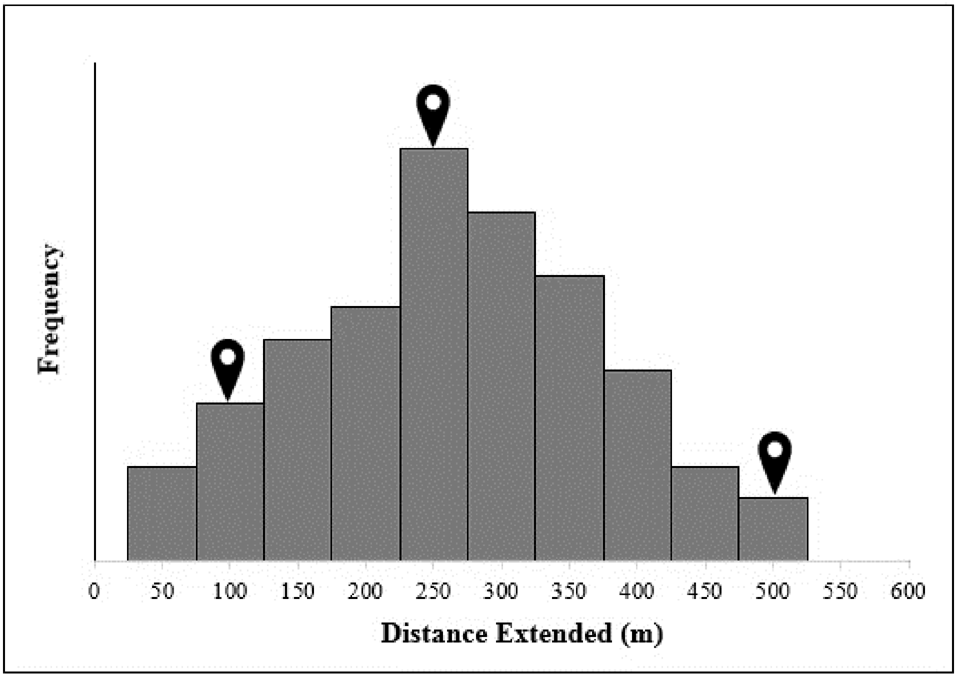


Figure S 4. Landslide – Multiple Salient Sites (350 m).

3. Battery Scenario Figures- Experiment 2

(Correct answer in brackets)

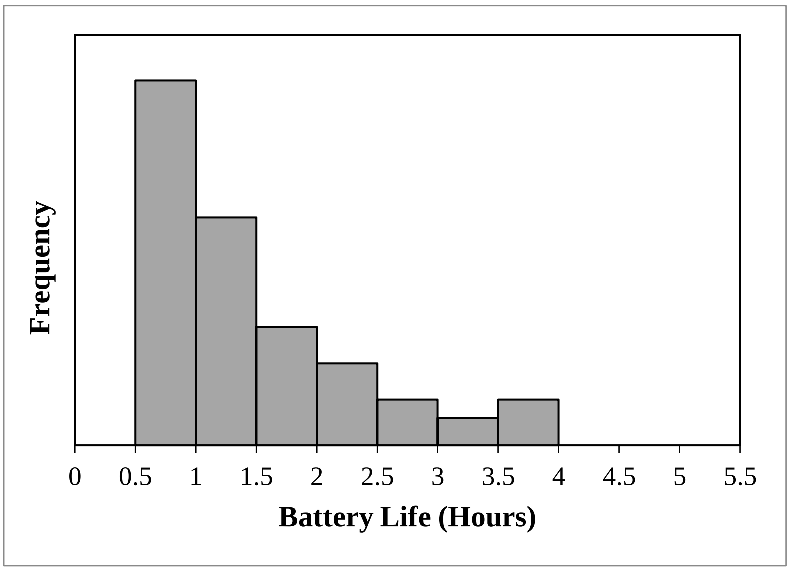


Figure S 5. Positive (2.25 hours).

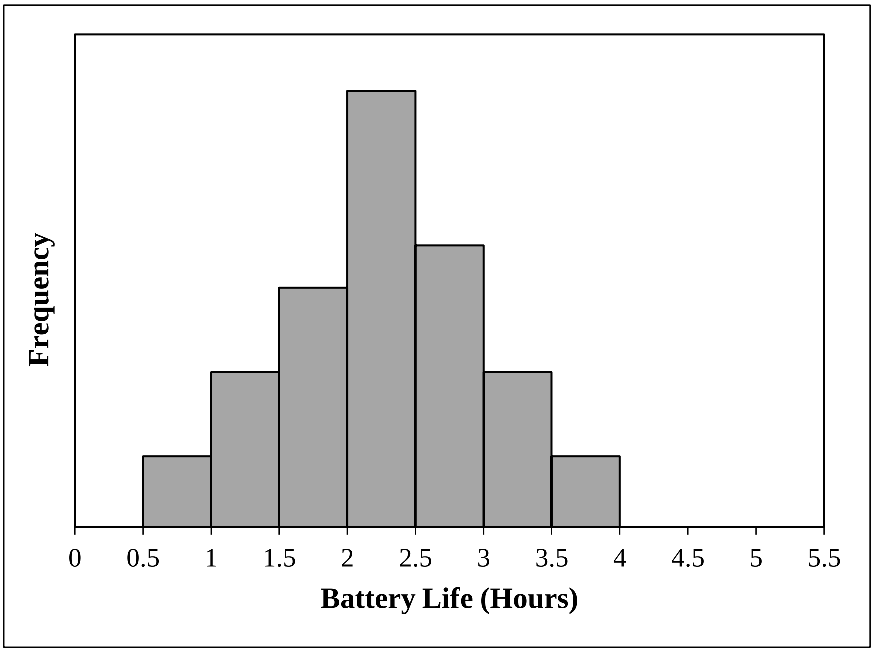


Figure S 6. Normal (2.75 hours).

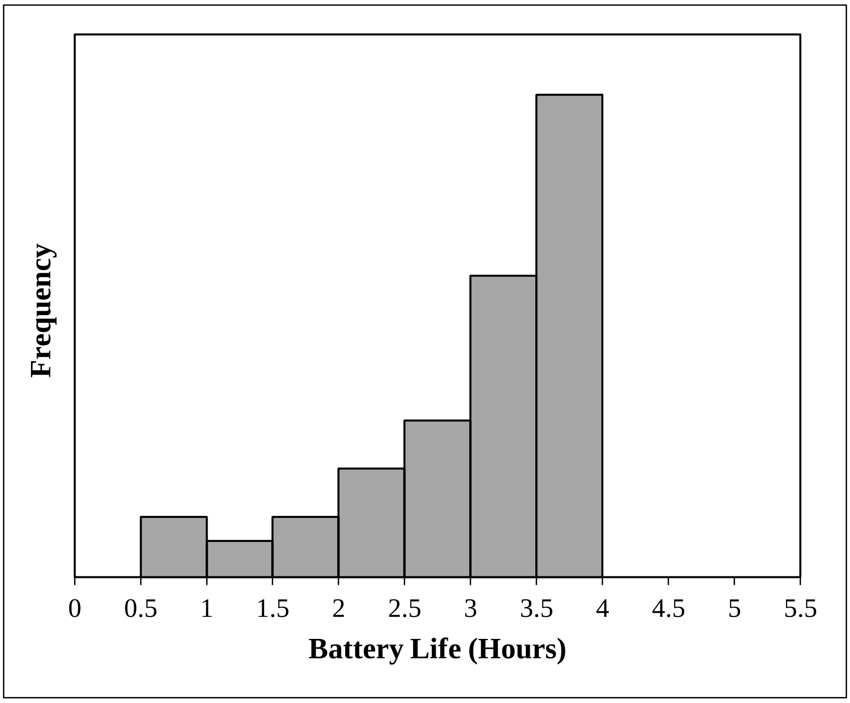


Figure S 7. Negative (3.75 hours).

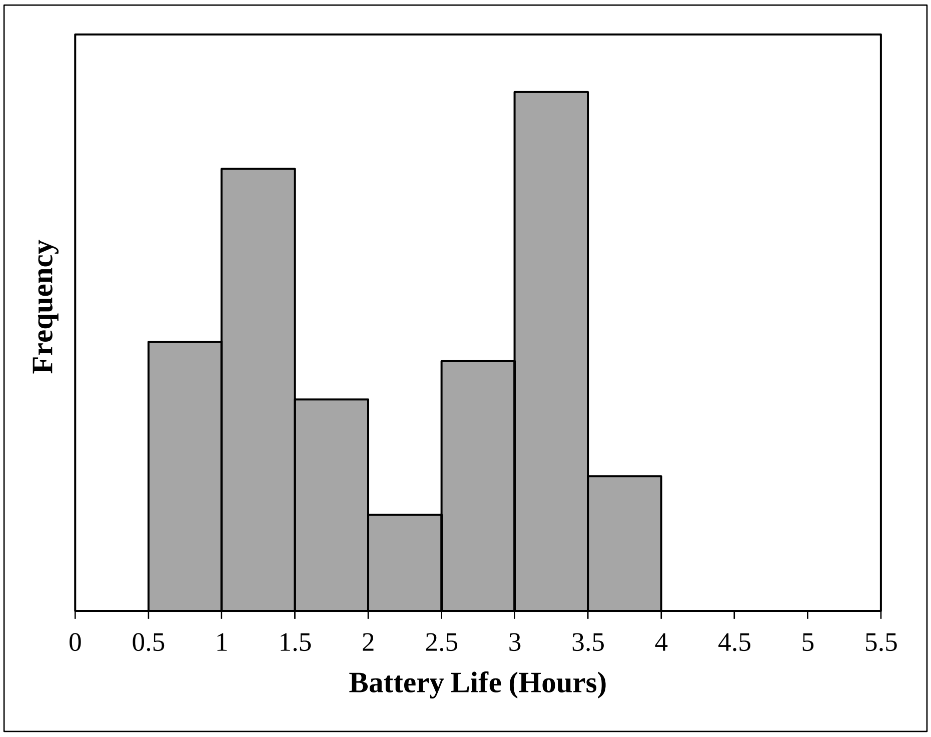


Figure S 8. Bi-Modal (3.25 hours).

4. Results – Cumulative Distribution Graphs – Experiment 2

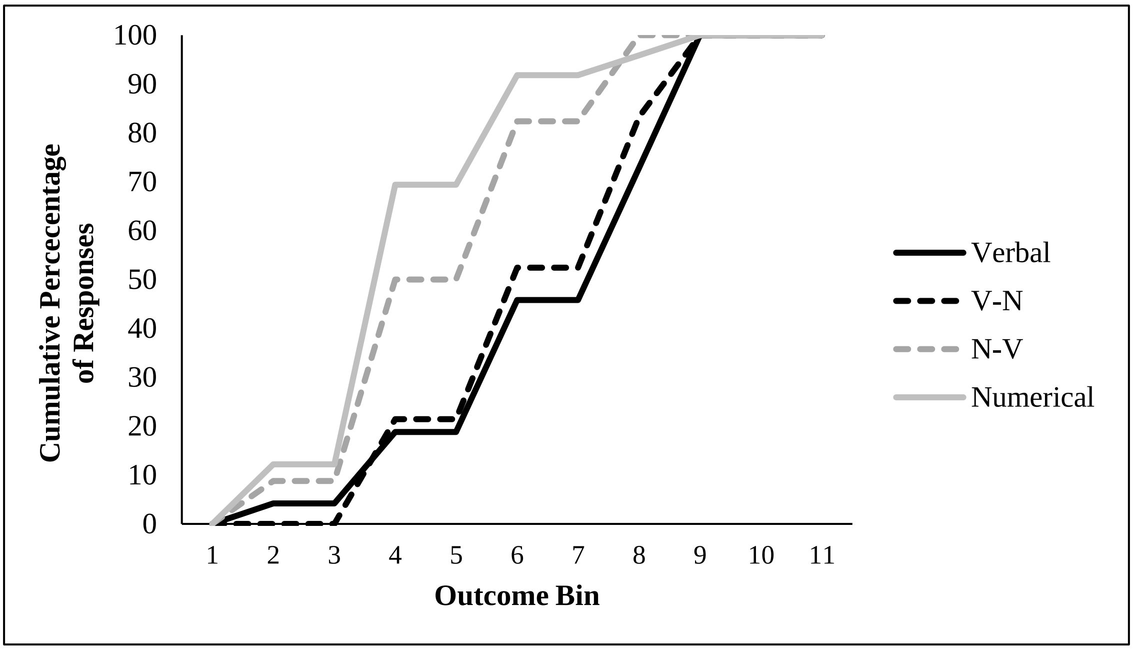


Figure S 9. Cumulative Distribution of Responses by Communication Format - Positive Distribution.

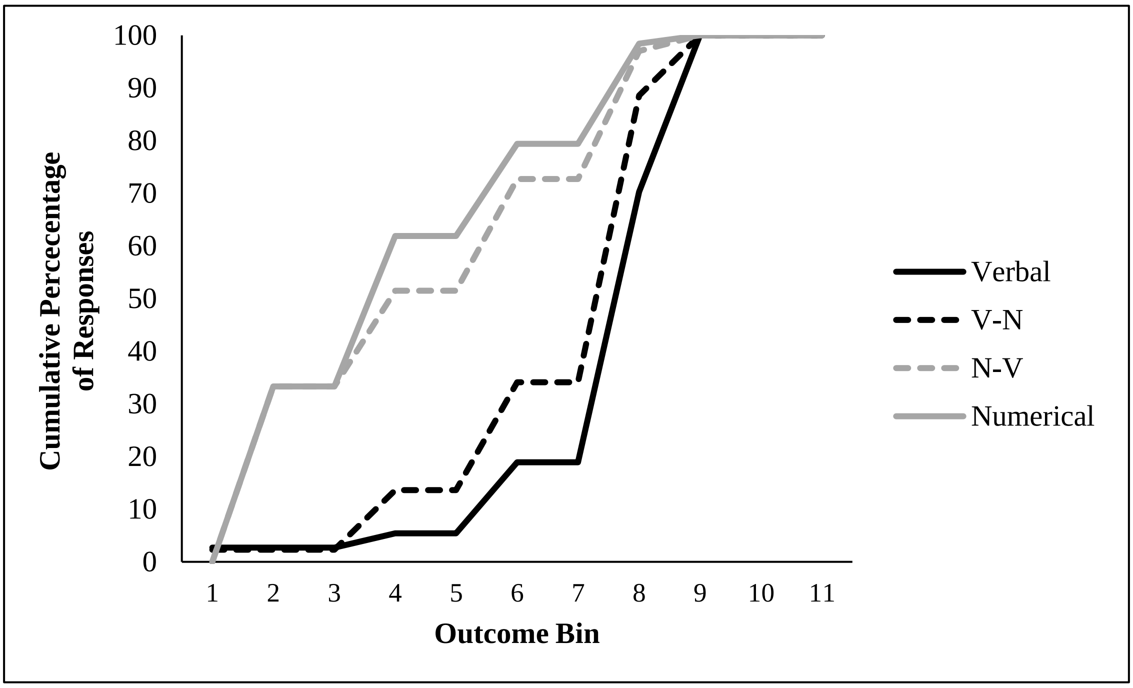


Figure S 10. Cumulative Distribution of Responses by Communication Format – Normal Distribution.

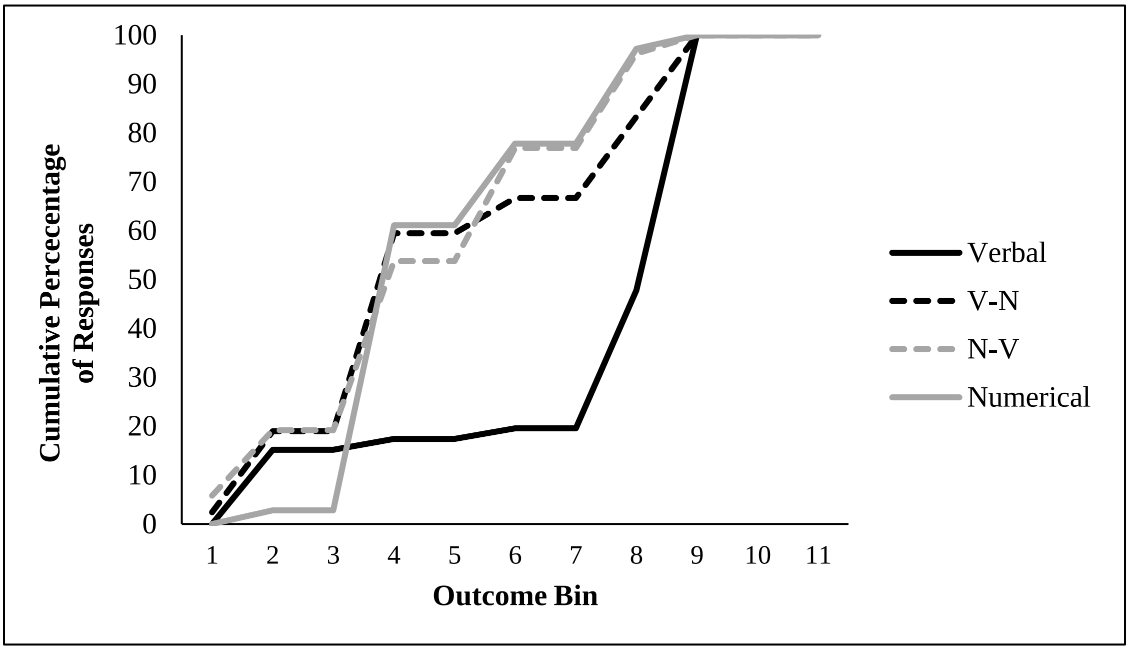


Figure S 11. Cumulative Distribution of Responses by Communication Format – Negative Distribution.

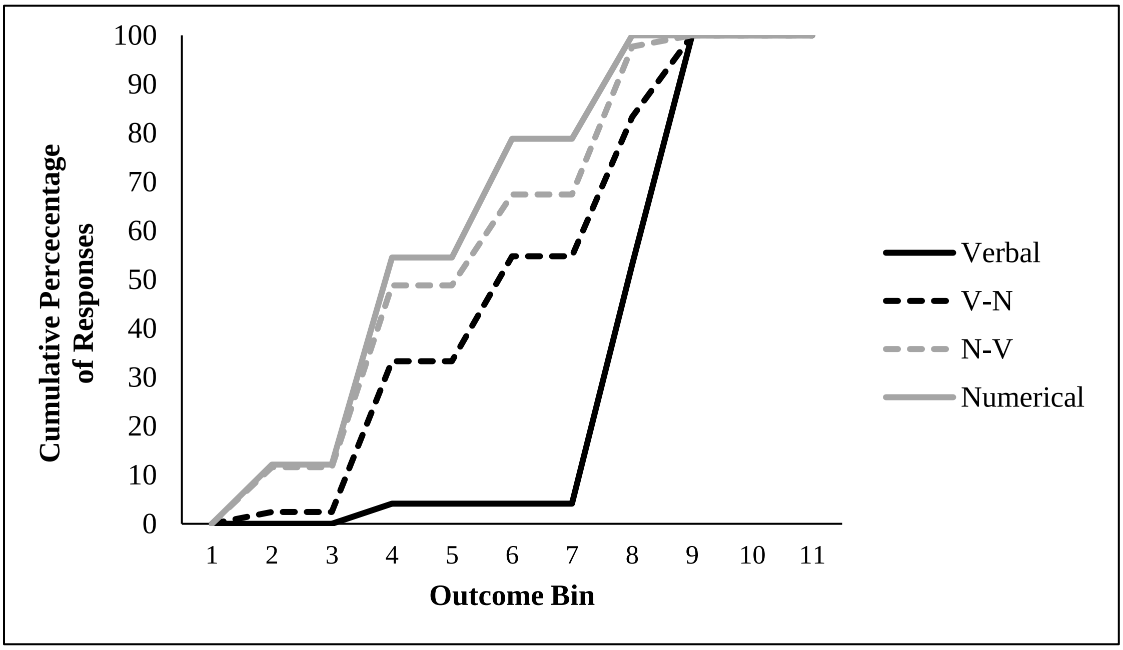


Figure S 12. Cumulative Distribution of Responses by Communication Format – Bi-Modal Distribution.

5. Experiment S1 – Clarifying the Mechanisms Behind Order Effects

In the context of the main experiments, it is not possible to make the distinction between the effect of parentheses and the effect of order. Walker (1823) suggests that in both writing and speech, parentheses are given little emphasis, with the option of either ignoring them, or pronouncing them with a lower tone. Subtly different to the parentheses account is Grice's (1975) conversational maxim that cooperative utterances will be 'orderly', with the most important information presented first. We therefore designed a further experiment to differentiate between the effect of parentheses and order.

Method

Participants

Three hundred participants were recruited for this online experiment via Amazon Mechanical Turk and were paid $0.20 upon completion. Eight cases were removed (due to duplicate IP addresses, for failing the attention check or completing the experiment quicker than reasonably expected) leaving a final sample of 292 (118 male) participants, aged between 18 – 73 (*Mdn* = 32).

Design, Procedure and Materials

Either brackets (B) or dashes (D) were used to create four different communication formats (V-N-B– "unlikely [10 – 30% likelihood]", N-V-B – "10 – 30% likelihood [unlikely]", V-N-D "unlikely – 10-30% likelihood" and N-V-D "10-30% likelihood – unlikely"), manipulated between-participants. Scenario (flood) was the same for all participants. As per Experiment 1, participants were required to type a numerical response which corresponded to the outcome being described. Finally participants completed a numeracy scale (Lipkus, Samsa, & Rimer, 2001), with two additional questions from the Berlin Numeracy Test (Cokely et al., 2012).

Results

Effect of Communication Format

The proportion of responses indicating high amplitude outcomes (the maximum / above maximum value present in the histogram) was highest in the V-N-B and V-N-D conditions. The N-V-B and N-V-D conditions had the lowest proportion of responses indicating high amplitude outcomes, χ^2 (3) = 13.62, p < .01. The distribution of responses reflected the order effects seen in the main paper, see Figure S 13.

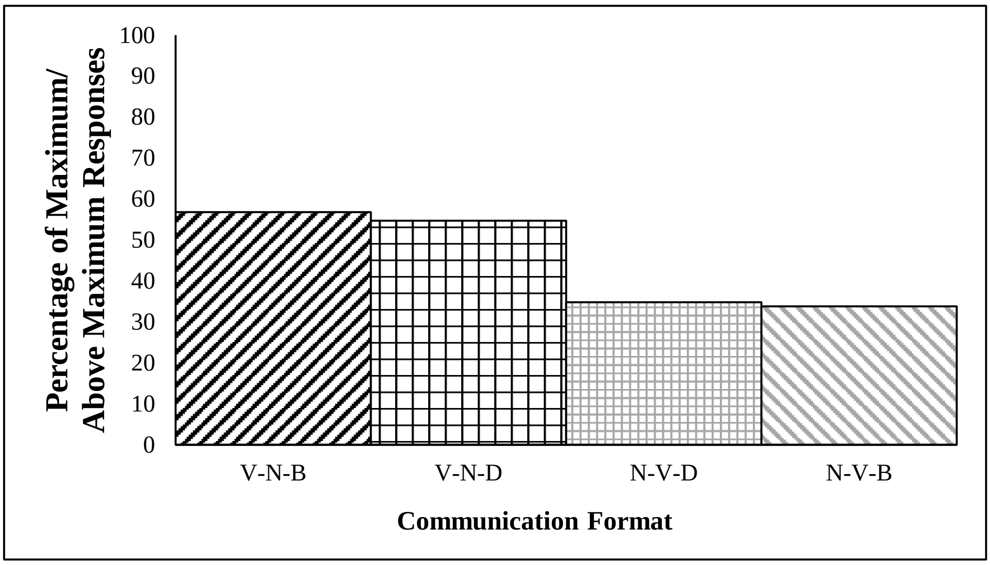


Figure S 13. Percentages of Maximum and Above Maximum Responses by Communication Format – Experiment S1.

A two-way ANOVA showed a significant effect of order F (1, 288) = 18.64, p < .001, η_p^2= .06 and no significant effect of punctuation, F (1, 288) = 0.40, p = .53. There was no interaction between order and punctuation, F (1, 288) = 0.15, p = .70. A post-hoc REGWQ procedure revealed that responses in the V-N-B (M= 7.32, SD= 2.99) and V-N-D (M= 7.41, SD= 2.89) were not significantly different. They were significantly higher than responses in the N-V-B (M= 5.64, SD= 3.26) and N-V-D conditions (M= 6.00, SD= 3.12). Responses were similar in the N-V-B and N-V-D conditions.

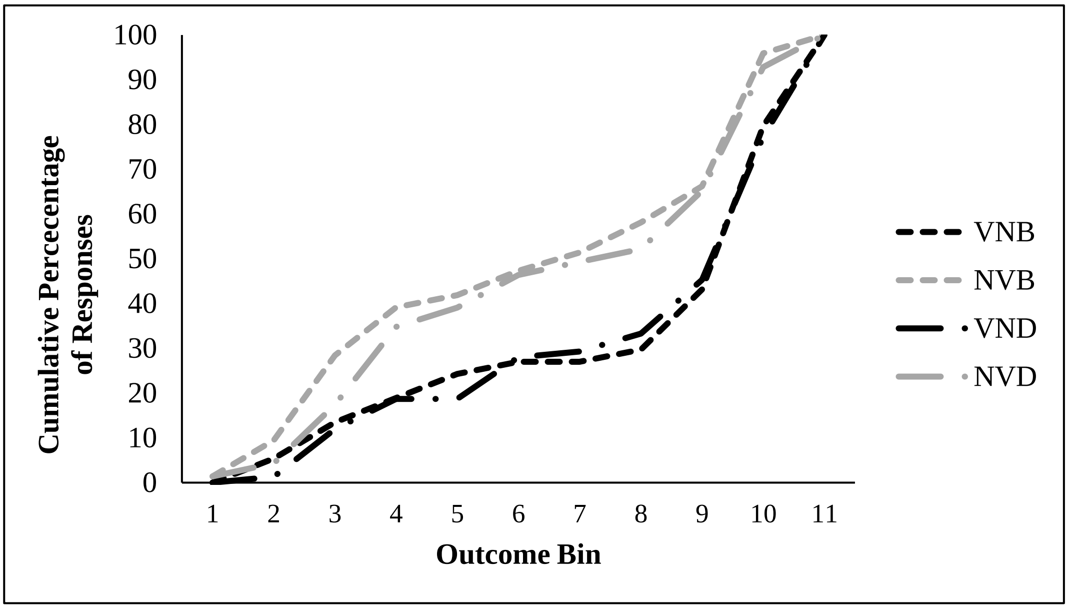


Figure *S 14* clearly displays the effect of order.

Figure S 14. Cumulative Distribution of Responses by Communication Format – Experiment S1.

As observed in the main paper, expressions where the numerical phrase came first had a higher proportion of answers below the correct answer (N-V-B: 51.4%, N-V-D: 49.3%) than above (N-V-B: 41.9%, N-V-D: 47.8%). Responses in the V-N-B and V-N-D conditions had a greater proportion of responses above the correct answer (V-N-B: 70.3%, V-N-D: 66.7%) compared to below the correct answer (V-N-B: 27.0%, V-N-D: 29.3%).

Effect of Numeracy

Answers for each question were coded as 1 if correct and 0 if incorrect, such that numeracy scores could range from 0 to 10. The distribution of numeracy scores is shown in Table S1.

Table S1. *Distribution of Numeracy Scores (%)*

	Score											
Numeracy Score	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Mean (SD)**
Expt S1	0.3	0.3	0.3	1.0	1.7	4.5	5.5	15.1	26.7	30.8	13.7	8.00 (1.64)

The effect of communication format held over differing levels of numeracy. A two-way ANOVA with numeracy (high / low) as a fixed factor showed there was a significant effect of communication format F (3, 284) = 5.83, p < .01, η_p^2= .06 and no significant effect of numeracy, F (1, 284) = 0.03, p = .86. There was no significant communication format × numeracy interaction, F (3, 284) = .58, p = .63.

There was a significant association between high and low numeracy and maximum / above maximum responses, χ^2 (1) = 5.34, p < .05, with those lower in numeracy giving a higher proportion of maximum / above maximum responses than those higher in numeracy.

6. Geological Scenario Graphs- Experiment 3

(Correct answer in brackets)

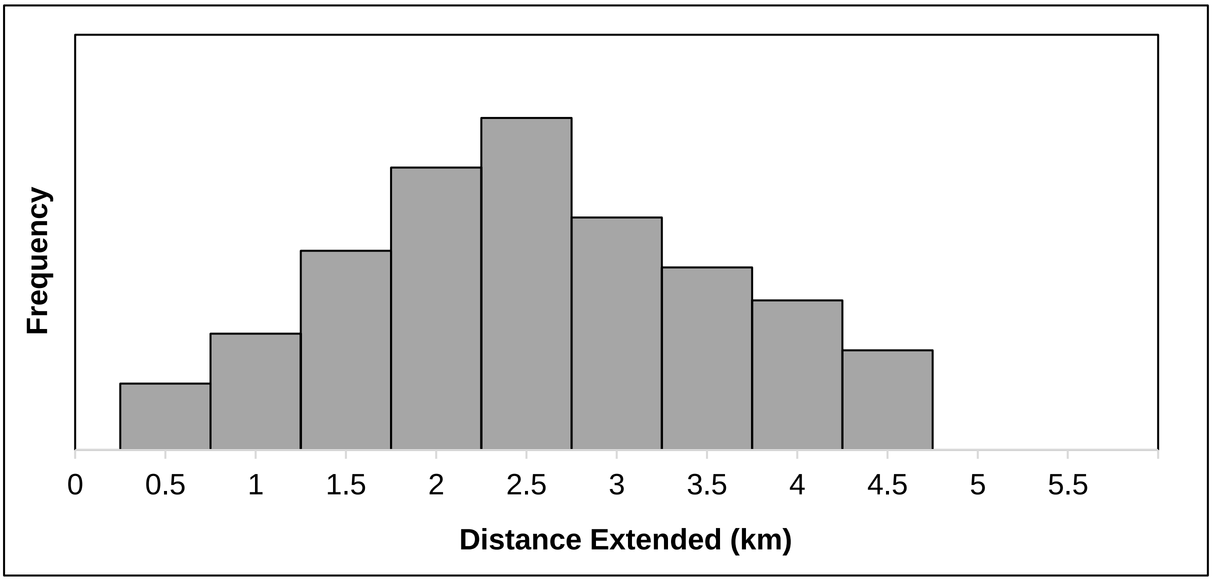


Figure S 15. Volcano (3.5 km).

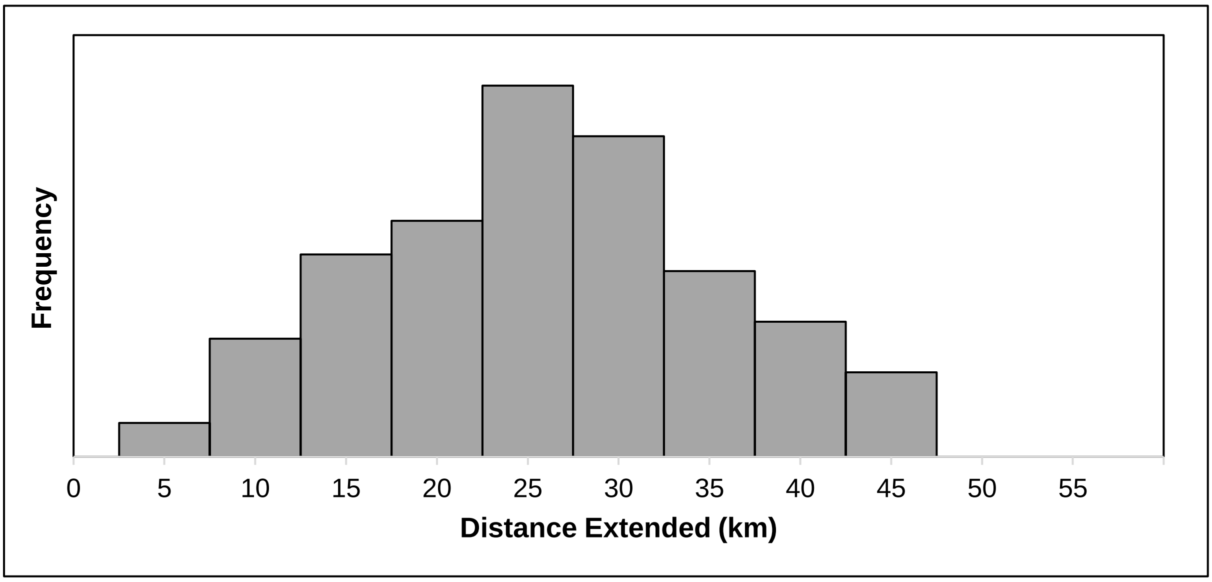


Figure S 16. Earthquake (35 km).

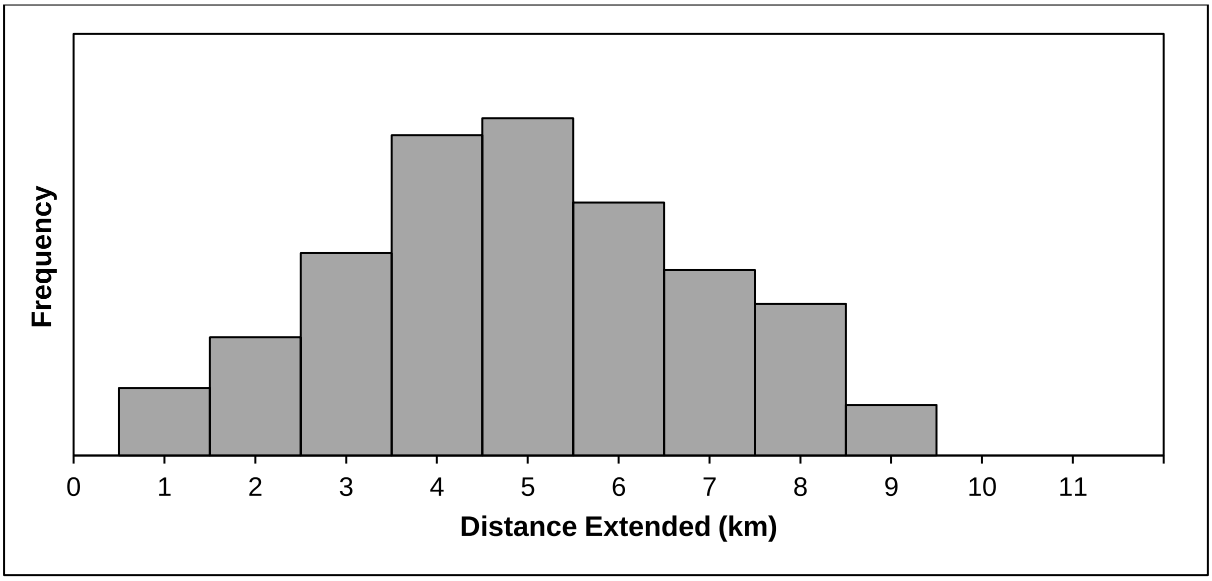


Figure S 17. Flood (7 km).

7. Experiment S2 – Extending Findings to Ranges

Method

Participants

Three hundred and twelve participants were recruited for this online experiment via Amazon Mechanical Turk and were paid $0.20 upon completion. 23 cases were removed (due to duplicate IP addresses, for failing the attention check or completing the experiment quicker than reasonably expected) leaving a final sample of 289 (114 male, 1 ‘other’) participants, aged between 19 – 70 (*Mdn* = 32).

Design, Procedure and Materials

Communication format (verbal – “unlikely”, numerical – “10 – 30% likelihood”, V-N – “unlikely [10 – 30% likelihood]” and N-V – “10 – 30% likelihood [unlikely]”) was manipulated between-participants. Scenario (flood) was the same for all participants. As per Experiment 1, participants were required to type a numerical response which corresponded to the outcome being described. Finally participants completed a numeracy scale (Lipkus et al., 2001), with two additional questions from the Berlin Numeracy Test (Cokely et al., 2012) included to increase variability in numeracy scores.

Results

Effect of Communication Format

The proportion of responses indicating high amplitude outcomes (the maximum / above maximum value present in the histogram) was highest in the verbal condition, followed by the mixed format condition. The numerical range condition had the lowest proportion of responses indicating high amplitude outcomes, χ^2 (2) = 76.35, $p < .001$.

The distribution of responses followed a very similar pattern to those in the main paper, see Figure S 18.

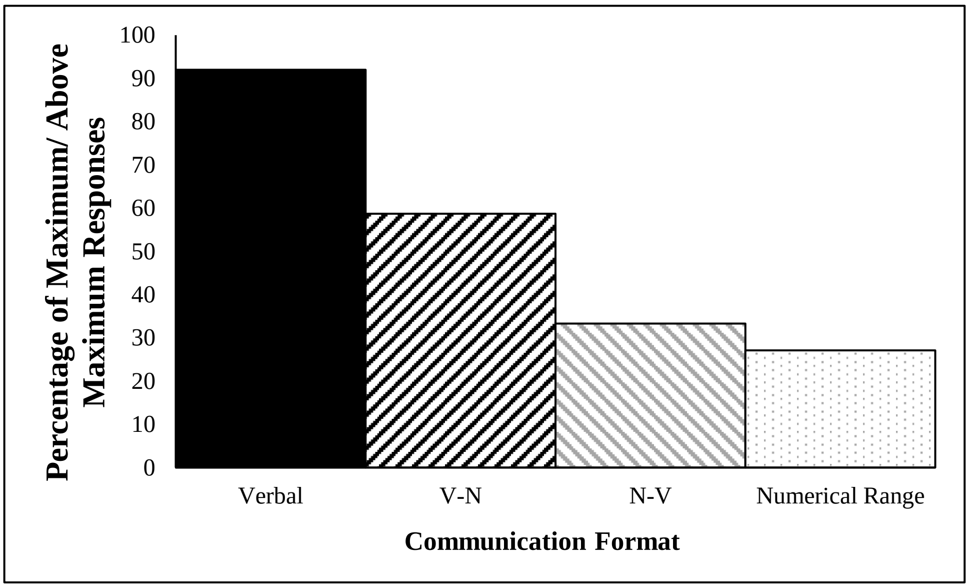


Figure S 18. Percentages of Maximum and Above Maximum Responses by Communication Format – Experiment S2.

A one-way ANOVA showed a significant effect of communication format F (3, 285) = 32.74, p < .001, η_p^2= .26. A post-hoc Ryan, Einot, Gabriel and Welsh Q (REGWQ) procedure revealed that responses in the verbal condition (M= 9.41, SD= 1.87) were significantly higher than responses in the other three conditions. Responses in the V-N condition (M= 7.60, SD= 2.81) were significantly higher than those in the numerical range condition (M= 5.47, SD= 3.05) and the N-V condition (M= 5.87, SD= 2.90). Responses were similar in the numerical range and N-V conditions. Again, there is clear evidence for the order effects mentioned in the main paper, see Figure S 19.

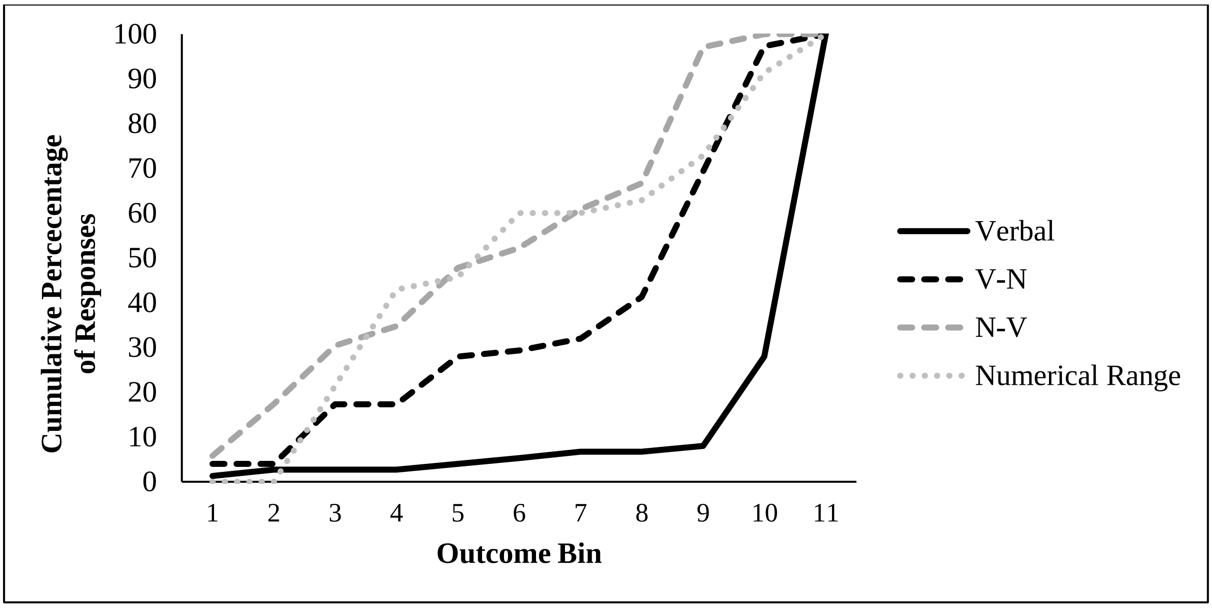


Figure S 19. Cumulative Distribution of Responses by Communication Format – Experiment S2.

As observed in the main paper, the numerical range condition had a higher proportion of answers below the correct answer (60%) than above (37.1%). To a lesser extent, the N-V condition followed this pattern, with fewer answers above the correct answer (39.1%) than below (52.2%). Responses in the V-N condition had a greater proportion of responses above the correct answer (68%) compared to below (29.3%).

Effect of Numeracy

Answers for each question were coded as 1 if correct and 0 if incorrect, such that numeracy scores could range from 0 to 10. The distribution of numeracy scores is shown in Table S2.

Table S2. *Distribution of Numeracy Scores (%)*

	Score											
Numeracy Score	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Mean (SD)**
Expt S2	0	2.1	0.3	1.7	2.4	3.1	5.5	13.1	29.4	31.8	10.4	7.84 (1.82)

The effect of communication format held over differing levels of numeracy. A two-way ANOVA with numeracy (high / low) as a fixed factor showed there was a significant effect of communication format *F* (3, 281) = 30.83, *p* < .001, η_p^2= .25 and no significant effect of numeracy, *F* (1, 281) = 2.67, *p* = .10, nor a significant format × numeracy interaction, *F* (3, 281) = 1.56, *p* = .20.

There was no significant association between high and low numeracy and maximum / above maximum responses, χ^2 (1) = 1.19, *p* = .29.