View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Explore Bristol Research

Donovan, A., Eiser, J. R., & Sparks, R. S. J. (2017). Expert opinion and probabilistic volcanic risk assessment. Journal of Risk Research, 20(6), 693-710. DOI: 10.1080/13669877.2015.1115425

Peer reviewed version

License (if available):
Unspecified

Link to published version (if available):
10.1080/13669877.2015.1115425

Link to publication record in Explore Bristol Research
PDF-document

This is the author accepted manuscript (AAM). The final published version (version of record) is available online via Taylor & Francis at http://www.tandfonline.com/doi/full/10.1080/13669877.2015.1115425. Please refer to any applicable terms of use of the publisher.

University of Bristol - Explore Bristol Research

General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available: http://www.bristol.ac.uk/pure/about/ebr-terms.html

Expert opinion and probabilistic volcanic risk assessment

Amy Donovan[1]

J. Richard Eiser[2]

R. Stephen J. Sparks[3]

[1] Department of Geography, University of Cambridge, Downing Place,
Cambridge CB2 3EN, UK. *ard31@cam.ac.uk (corresponding author)*

[2] Department of Psychology, University of Sheffield, Western Bank,
Sheffield S10 2TP, UK. *.j.r.eiser@shef.ac.uk*

[3] Department of Earth Sciences, University of Bristol, Queen's Road,
Bristol BS8 1RJ, UK. *steve.sparks@bristol.ac.uk*

Abstract

We present data from an international survey of scientists working at volcanic observatories concerning eruption likelihoods. The scientists were asked a range of questions using different types of phrasing. The data suggest that the phrasing of questions affects the ways in which probabilities are estimated. In total, 71% of respondents (N=70) exhibited some form of inconsistency in their answers between and/or within different question formats. The data also allow for an analysis of the use of scaling in probabilistic assessment, and the use of quantitative versus verbal risk measurements. However, some respondents were uncomfortable with providing any numerical probability estimate, perhaps suggesting that they considered the uncertainty too high for meaningful judgements to be made.

Keywords: expert judgement; volcanic risk; probability

1 INTRODUCTION

Research in psychology on how individuals estimate probabilities (Tversky and Kahneman, 1973,1974; Lichtenstein and Fischoff, 1980; Gigerenzer, 2008; Kynn, 2008) indicates that there are widespread issues in general populations with estimating probability that commonly results in inconsistencies, inaccuracies and biases. With the increasing use of expert judgement in forecasting of natural hazards and assessment of risk, this paper examines whether these issues are evident in groups of scientists[1]. We focus in particular on volcanologists who are working in volcano monitoring institutions and officially provide information for governments during volcanic crises. Training is very_m variable among volcanologists in statistics and probabilities. So although a technical group might be expected to have a better understanding of probabilities, there nonetheless may be deficiencies in education or other factors that lead to poor understanding of probabilities and similar problems to the wider population. Related to these issues is the matter of communicating probabilistic forecasts to a general public and decision-makers who do not necessarily have a technical background. Qualitative descriptors of probabilities, such as “very likely” or “unlikely” might be used to convey the forecast. Here a question arises as to whether numerical values of probability match qualitative descriptors consistently. In this paper, we survey 95 volcanologists, examining in particular responses relating to the probability of an eruption from the volcano that they were working on (specified

[1] In this paper, “scientists” and “experts” refer to the volcanologists surveyed. The terms are used interchangeably.

individually at the start of the survey). We examine the probability distributions that were produced, the ways in which different probabilities relate to one another, and some inconsistencies between the responses to some of the questions.

2 THEORETICAL BACKGROUND

Volcanic risk is highly uncertain. Many of the signals that are measureable at the surface are indicators that magma is moving, but do not guarantee that it will erupt. Forecasting volcanic eruptions is therefore challenging (e.g. Sparks, 2003; Cashman and Sparks, 2013) – but scientists are frequently asked for assessments because of the high stakes involved when a population is in danger. In many countries, volcano monitoring (where it is done at all) is the responsibility of particular agencies, often called "volcano observatories" (e.g. Donovan and Oppenheimer, 2015). The day-to-day work of volcano observatories involves the collection and analysis of a range of datasets, usually focussed on measuring seismicity, ground deformation and volcanic gas emissions (e.g. Tilling, 2008). Staff at volcano observatories includes scientists and technical specialists, who may or may not have postgraduate qualifications. In referring to these scientists as "experts", we take the view that formal qualification is not necessarily an indicator of expertise (e.g. Wynne, 1996). Volcano observatory scientists are therefore referred to as experts, because they study the activity of volcano on a daily basis and are the official source of information about the volcano to civil protection organisations and governments.

The high levels of uncertainty associated with volcanic emergencies are well documented (Aspinall et al., 2003; Newhall and Hoblitt, 2002; Marzocchi et al., 2012; Marzocchi and Woo, 2009; Sparks et al., 2013). The assessment of risk from volcanoes requires the compilation of multiple strands of evidence, datasets, models and methods (Aspinall et al., 2003). In order to make recommendations based on diverse forms of evidence, belief-based probabilistic methods such as expert elicitation and Bayesian methods (e.g. Aspinall et al., 2002; Aspinall, 2006; Hincks et al., 2014; Marzocchi et al., 2004, 2008; Neri et al., 2008; Donovan et al., 2012) are increasingly being advocated and applied. These methods have been effectively used on Montserrat over a long period (Aspinall et al., 2002; Donovan et al., 2012). Other forms of structured and unstructured probability estimation may also be used in an emergency. Yet questions remain about the ways in which experts perceive probabilities and respond to questions with different formulations (e.g. O'Hagan et al., 2006; Doyle et al., 2014). This study explores volcanologists' judgements of probabilities in an informal and theoretical context (as opposed to an operational setting), in order to assess the heuristics, role of framing and role of scaling in their responses. This section reviews some of the literature on probability judgements.

2.1 Measures of judgment

Individuals vary in their ability to judge probabilities (e.g. O'Hagan et al., 2006; Tversky and Kahneman 1983; Kahneman and Tversky, 1982; Gigerenzer et al., 2005, 2007), possibly associated with reliance on heuristics (Tversky and Kahneman, 1974), the interpretation of which may depend on views about the nature of probability itself (e.g. Kahneman and Tversky, 1996; Gigerenzer, 1994; Vranas, 2000). In this section, we contextualise the following discussion in terms of how the accuracy of probability judgements may be measured, looking in turn at calibration, informativeness and reliability. Calibration is a measure of the relationship between a set of subjective probability judgements and the corresponding relative frequency of the events in question. It is affected by over/underconfidence, over-extremity and discrimination (see O'Hagan et al., 2006 for a detailed review). Overconfidence is defined as the view that an event is more likely than the relative frequency; over-extremity is the tendency to overestimate low probabilities and underestimate high probabilities; discrimination is the ability to discriminate between low and high probability events. There is some evidence that calibration of experts can be improved with training (e.g. with feedback; Ferrell, 1994; Fischoff, 1991). It may also be context-specific – there is evidence from some fields that the potential perceived impact of a particular event may affect judgements (e.g. see discussion in O'Hagan et al., 2006). In volcanology, calibration is difficult to determine empirically, because most of the events in which volcanologists are interested are single events without frequency data – such as the likelihood of an explosion on a particular volcano in the next six months. There are, however, other methods that can estimate calibration (e.g. Cooke 1991).

Calibration describes the accuracy of a judgement (whether or not it is correct); informativeness describes the precision of a judgement (how well resolved it is around the correct value). It refers to the uncertainty bounds on an expert's response to a probability question, and therefore also reflects their confidence in their own estimate. Calibration and informativeness have been shown to be distinct from the level of disciplinary expertise that an individual may have: tests of knowledge do not correlate with calibration (e.g. Kahneman et al., 1982), although in general experts may produce more convincing responses than lay people (e.g. Murphy et al., 1984). However, calibration is not the only issue involved in probability judgements. There are also broader questions about how experts interpret probability questions and probability theory: how consistent they are when different wordings are used, for example. We turn now to discuss the social psychology literature that seeks to explain the ways in which such judgements are made.

2.2 Heuristics and biases

Probability estimation by both experts and laypeople has been investigated within psychology (Tversky and Kahneman, 1973, 1974; Lichtenstein and Fischoff, 1982; Kynn, 2008; Kahneman et al., 1982; Kahneman and Tversky 1972; Plous, 1989; Gigerenzer, 2008; Goldstein and Rathschild, 2014; Slovic, 2000). These studies have suggested that the estimation of probabilities relates to a series of heuristics and biases. Heuristics are rules that are applied in the decision to place a particular probability on an event: they are difficult to quantify and have varying degrees of effect in different individuals. In probability estimation, they may include "rules of thumb" about the reliability of certain types of data or the relative importance of types of data. Understanding the variability in the use of heuristics provides insights into the cognitive processes involved in assigning probabilities, which is important in risk assessment and communication.

We consider here three examples of heuristics that are relevant to the estimation of probabilities: base-rate neglect, anchoring and availability. Base-rate neglect occurs when individuals partly neglect base-rate frequencies, while relying on cues or informal decision rules ('heuristics'), which may have dubious validity. The problem with base-rate neglect is that it is not meaningful to have two probabilities for the same event, one that is based solely on frequency and one that incorporates new information such as volcano monitoring data. For example, the probability of an eruption in the next year at a particular volcano is 0.2 based on historical frequency. If a seismic swarm occurs, however, we may consider that the probability of an eruption in the next year has increased above the base-rate. Gigerenzer (1991) has emphasised the distinction between single-event probabilities and frequencies as a subset of the belief-based–frequentist divide. He notes that "Probability theory is about frequencies, not about single events. To compare the two means comparing apples with oranges" (1991: 88).

This comment sums up the need for belief-based methods: they allow the evaluation of the probability of individual events based on a wide range of evidence. The comparison of probabilities obtained from frequency analysis and those obtained from belief-based analysis is not conceptually meaningful (Dawid, 1982; Hacking, 2001) for a single event (though belief-based analysis may take frequencies into account). The relationship between frequency-based and belief-based probabilities is not straightforward, and depends upon individual philosophies of probability: “strong frequentists” argue, with (Gigerenzer, 1991), that only frequencies are valid and that single-event probabilities are meaningless and vulnerable to heuristics such as base-rate neglect, while Bayesians argue that frequencies are limited in their usefulness for real-world problems and decision-making under uncertainty.

In addition to base-rate neglect, the availability bias occurs when subjects base their estimation on previous experience – something that is “available” to them in their memory (Tversky and Kahneman, 1973). Anchoring occurs when individuals base their estimates on scaling up or down from a “set” value that they have decided upon, and may result in systematically imposed probability distributions based on scaling. The use of heuristics provides some explanation of the variability in belief-based probability estimates, and has been applied in a slightly different way to issues around expert judgement (MacGillivray 2014).

There have been a number of studies dealing with potential differences in heuristics between “lay” and “expert” groups. Many of these studies suggest that experts are susceptible to the same heuristics and biases as laypeople: O’Hagan et al., 2006 state that “substantive expertise in a specialist area is no guarantee of normative expertise in providing coherent probability assessment” (p.58; see also for example reviews in Doyle et al., 2014; Smith and Kida, 1991;

Burgman et al., 2011). A distinction may be drawn between expertise in a particular discipline (such as volcanology), and expertise in the provision of probability estimates (e.g. Doyle et al., 2014; Cooke, 1991; Burgman et al, 2011; Martin et al., 2012). Since probability estimation is increasingly used to inform volcanic risk assessment, it is important to understand the variability of experts' abilities to provide uncertain judgements, and any factors that affect that variability.

2.3 Issues of framing in probability estimation

There are several ways in which probability judgement may be affected by the wording of the question that is asked and how the question is therefore conceptually represented, known as "framing" (e.g. O'Hagan et al., 2006; Gigerenzer, 1991; Hoffage et al., 2000).
Some studies have examined in particular the language that is used in framing probability estimates (e.g. Doyle et al., 2014; Brun and Teigen, 1988; Teigen and Brun, 1999, 2003; Risbey and Kandlikar, 2007; Doyle et al., 2011). Positive and negative framing, for example, may impact the response to probability questions (Teigen and Brun, 1999). Probability questions can also be phrased to help or hinder respondents in the avoidance of typical "traps" (Teigen and Brun, 1999).

In addition to issues of language, the framing of questions in terms of time is important in determining probability judgements. Questions can be expressed as a likelihood of an event in particular time period, or as the time period in which the event has a particular likelihood of occurring, for example. People may be shy of giving a probabilistic frequency estimate of "1 in 1" (ie, with stating that an event will certainly occur), but be more content with estimating a number of days or years within which the event will occur. The use of different time periods that show some kind of scaling (in this paper, we use 3, 30 and 300 years) can also have an impact on the heuristics applied by individuals. Doyle et al (2011, 2014) explored some of the implications of time frames on estimates, and showed that estimates were biased over the longer time periods due to anchoring in the shorter periods.

2.4 Verbal versus numerical risks

There are two primary ways of providing forecasts – numerical and verbal. Verbal assessments may involve a verbal scale, such as “very high chance” to “very low chance” or even “no chance at all”. Individual judgements about the use of such scales have been shown to depend on their perception of the scale itself (Parducci, 1965; Eiser and White, 1974). Verbal scales cannot be related easily to numerical estimates, since the perception that something is a “high” risk may depend on subjective, social factors. People react to the wording of the scale (Eiser and Hoepfner, 1991) and to the perceived consequences of the risk being realised (Teigen and Brun, 2003; Bruine de Bruin et al.2000; Karelitz and Budescu, 2004). The interpersonal variability in reactions to a verbal scale is a significant issue for risk communication, since the wording of risk assessments affects their impact on policymakers and the public.

In this paper, we use different phrasings of questions with regard to time period to investigate the role of scaling and other heuristics in the estimation of probabilities of eruption.

3. METHODS

3.1 Survey design

A survey of scientists working at volcano observatories was carried out in 2011. The use of a quantitative survey was judged the best method to obtain a statistically useful sample of scientists from a large number of volcanoes. While the size of the sample allows only explorative statistics, it provides valuable insights. Scientists were asked a range of questions about a volcano that they work on – which they specified early in the survey. In this paper, we report the results from the questions regarding the likelihood of eruptions at 'their' volcano within different time-frames, using different question formats. Specifically, these response formats involved (a) *likelihood ratings* – "How *likely* is it that [volcano name] will have a major eruption within the next [N] years?", with responses on a 7-point scale from 'extremely unlikely (1) to 'extremely likely' (7); (b) *chance estimates* – "Please express your estimate in terms of "1 chance in..." (E.g. "1 chance in 10", "1 chance in 1000" – fill in ANY number you wish. If you are absolutely certain an eruption will occur, say, "1 chance in 1")" ; these pairs of questions were repeated for N = 3, 30 and 300 years; and (c) *timescale forecasts* – "I believe that [volcano name] will certainly have a major eruption some time within.... years" and "I believe that there is at least a 50:50 chance that [volcano name] will have a major eruption within the next...years", respondents being required to fill in the number of years (d) *confidence* – "How confident are you in the estimates you have given above (taken as a whole)?" (extremely unconfident (1) to extremely confident (7)). Additional responses from the survey are published elsewhere (Donovan et al., 2014).

In addition to these questions, respondents were also asked for their age, nationality, gender, highest level of education and experience working in their current role. These variables were used as predictor variables in the statistical analyses in order to assess demographic factors that might affect the result. Unless stated, demographic variables had no significant effect on the responses.

The questions are summarised in Table 1.

	Question	Response type
S3	How *likely* is it that [volcano name] will have a major eruption within the next 3 years?	Scale from 'extremely unlikely (1) to 'extremely likely' (7)
C3	Please express your estimate in terms of "1 chance in..." (E.g. "1 chance in 10", "1 chance in 1000" – fill in ANY number you wish. If you are absolutely certain an eruption will occur, say, "1 chance in 1")	Chance estimate
S30	How *likely* is it that [volcano name] will have a major eruption within the next 30 years?	Scale from 'extremely unlikely (1) to 'extremely likely' (7)
C30	Please express your estimate in terms of "1 chance in..." (E.g. "1 chance in 10", "1 chance in 1000" – fill in ANY number you wish. If you are absolutely certain an eruption will occur, say, "1 chance in 1")	Chance estimate
S300	How *likely* is it that [volcano name] will have a major eruption within the next 300 years?	Scale from 'extremely unlikely (1) to 'extremely likely' (7)
C300	Please express your estimate in terms of "1 chance in..." (E.g. "1 chance in 10", "1 chance in 1000" – fill in ANY number you wish. If you are absolutely certain an eruption will occur, say, "1 chance in 1")	Chance estimate
T1	I believe that [volcano name] will certainly have a major eruption some time within.... years	Time estimate

T50	I believe that there is at least a 50:50 chance that [volcano name] will have a major eruption within the next...years	Time estimate
Con	How confident are you in the estimates you have given above (taken as a whole)?	Scale from extremely unconfident (1) to extremely confident (7)

Table 1. Questions applied in the survey. Note that scale variables only gave verbal descriptions for the two ends of the scale. The codes provided in the first column are used later to demonstrate how the derived scores were calculated. “S” variables are those referring to the verbal scale; “C” variables are chance estimates; “T” variables are referred to as timescale estimates

3.2 Statistical methods

Initially, several derived scores were calculated. Three sets of transformations were carried out on the chance estimates to assess the use of anchoring and test for any inconsistencies in experts’ answers: C30 was divided by C3; C300 was divided by C30; and C300 was also divided by C3.T50 was also divided by T1 variable for a similar reason (see Table 2). Dummy variables were then created to indicate whether or not scaling by an integer was used in calculating the answers. The S variables (i.e. verbal scale estimates, see Table 1) were compared with the chance estimates (C values) to look at how the verbalscale was used in relation to probability estimates. Four other dummy variables were also calculated to look for anchoring effects and inconsistencies between different timescales and framings. These checked for (i) consistency between the 3 year and 30 year chance estimates (C3 and C30); (ii) consistency between the 30 and 300 year chance estimates (C30 and C300); (iii) consistency between chance variables and “certainly within” (C values and T1); (iv) consistency between chance variables and 50:50 (C values and T50); (v) consistency between the two timescale values (T1 and T50). Finally, a variable representing “any inconsistency” was calculated (Con). Inconsistencies involved:

- Higher probabilities for eruptions over shorter timescales than longer ones;
- Longer timescales for a 50:50 chance of eruption than for certainty of an eruption;
- Discrepancies between the chance distributions and the timescales, such as a 1 in 10 probability of an eruption in the next 300 years, but certainty in an eruption in 200 years.

The calculated variables are summarised in Table 2.

The variables were tested using a range of statistical methods. As the dataset failed tests for normality and homogeneity of variance, non-parametric tests were applied. To examine the relationship between scale variables, Spearman’s ρ was used. To assess the relationships between nominal and scale variables, the Kruskal-Wallis analysis of variance was used (H). This is a non-parametric test that ranks the median values within each category of the predictor. Similarly, for two-category variables, the Mann-Whitney test was used (U). Neither of these tests produced significant results. Finally, relationships between categorical variables were examined using Pearson’s $\chi^2$and the likelihood ratio. All tests were assessed at the 5% significance level.

Dummy variable	**Calculation**	**Meaning**
Check3-30	C30/C3	If >1, then the values are consistent (ie, an

		eruption in 3 years is less likely than in 30 years)
Check30-300	C300/C30	If >1, then the values are consistent (ie, an eruption in 30 years is less likely than in 300 years)
Check3-300	C300/C3	If >1, then the values are consistent (ie, an eruption in 3 years is less likely than in 300 years)
Checktime	T50/T1	If <1, then the values are consistent (ie, the time period given for a 50-50 chance of eruption is shorter than that for “certainly within”.
Consist3_30	Coded based on Check3-30	0=inconsistent, 1=consistent (ie value for 30 years greater than or equal to 3 years)
Consist30_300	Coded based on Check30-300	0=inconsistent, 1=consistent (ie value for

		300 years greater than or equal to 30 years)
Consist_CT1	Coded based on T1 and C3, C30, C300	0=inconsistent, 1=consistent (i.e., if T1<300 and C300>1, then answer is inconsistent;
Consist_CT50	Coded based on T50 and C3, C30, C300;	0=inconsistent, 1=consistent (i.e. the probabilities given in for C3, C30 and C300 are consistent with the time period given for T50.)
AnyInconsist	If any of consistency variables =0, then 0; else 1.	Is there any inconsistency for this respondent?

Table 2. Calculations carried out on the variables in Table 1 to produce indicators of consistency in estimates.

4 RESULTS

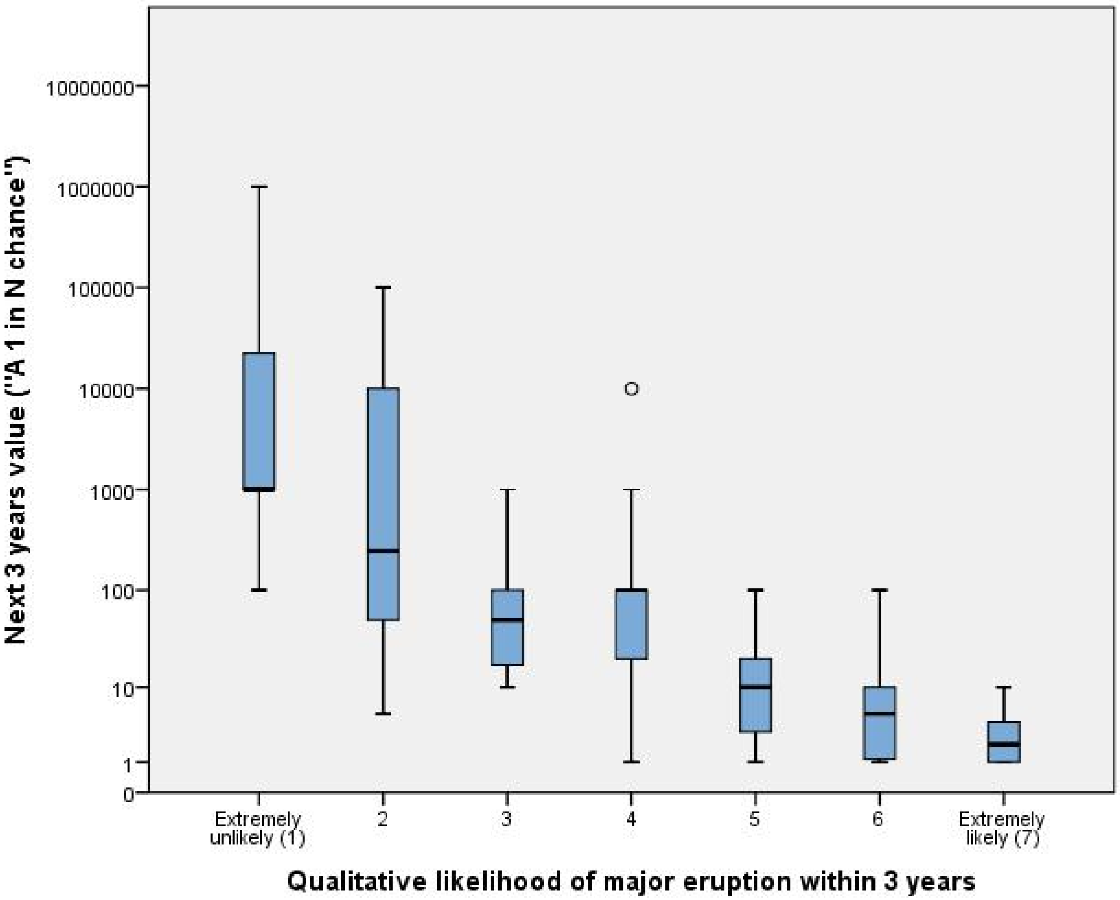
4.1 Survey demographic

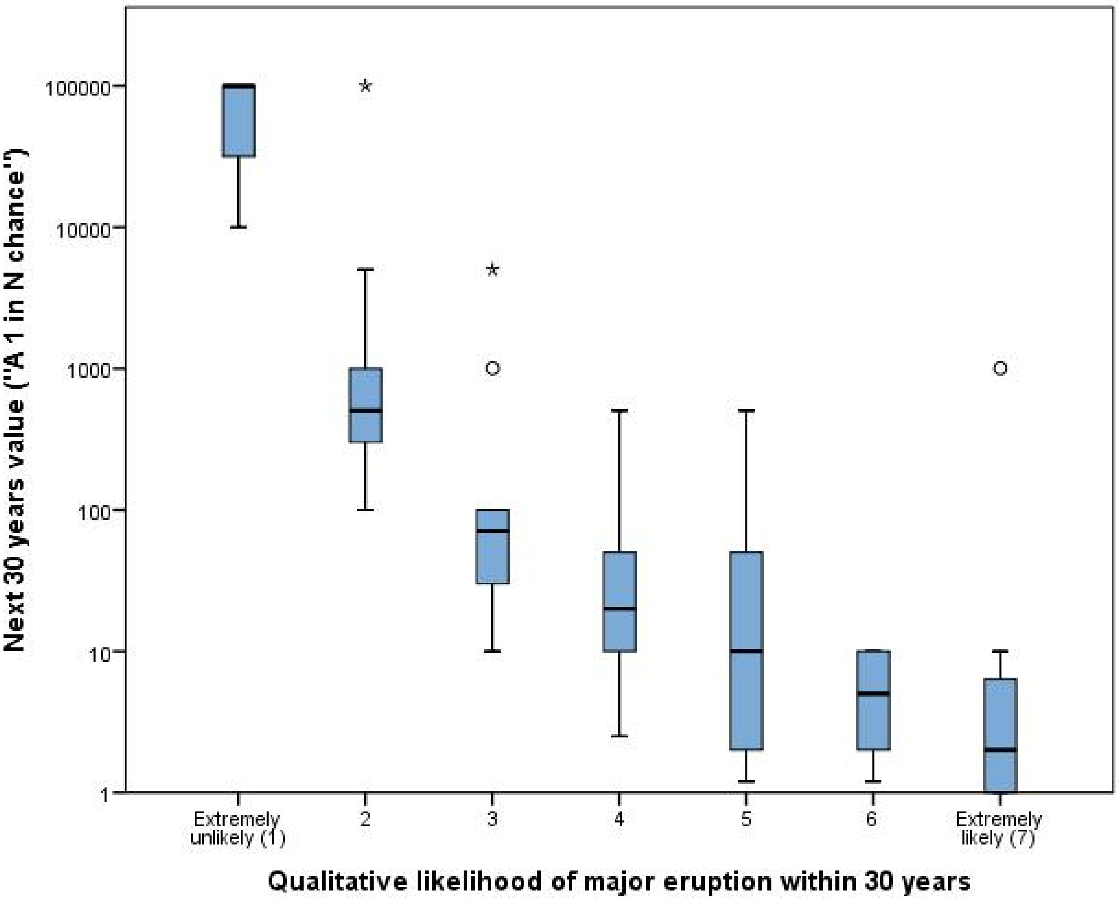
The scientists who responded to the survey varied in their willingness to answer all of the questions – particularly those questions that required numerical responses. In total, 111 scientists started the survey, but only 95 answered the questions considered in this paper. In the case of chance estimates only 74 responded out of the 95. Of these 74 respondents, 20 were women and 54 were men. In terms of tectonic setting, 12 worked on volcanoes associated with mantle plumes and 59 at volcanoes associated with subduction zones (3 failed to specify their volcano). In total, 45 had doctoral degrees and 16 had Masters degrees (volcano observatory scientists do not always have higher degrees). Fifty had lived in their current community for more than 10 years, and 44 lived within 120 km of the volcano in question. The distribution of ages was bimodal (with modes at ~30 and ~50), with a mean of 46. The distribution of the larger dataset was similar in structure with a mean at 44. Nine scientists out of the 95 had responded to the “half-half” and “certainly within” variables, but not the chance estimates.

4.2 Verbal and numerical scales

Of the 21 scientists unwilling to provide chance estimates, some stated that this was due to the high levels of uncertainty involved. Men were 3 times more likely to give quantitative estimates than women (based on the odds ratio; $\chi^2 = 4.8$, $p<0.05$), but no other demographic variables were significant. The following sections refer to the 74 scientists who provided responses to both timescale and chance ratings, with the exception of 4.5, which refers to the 83 who gave timeframe estimates.

Figure 1 shows the relationship between the probability value given, and the use of the verbal scale. This demonstrates that over both short and long timescales, lower probabilities are more difficult to estimate and the relationship between the perception of a “low” probability and a “high” probability on a verbal scale varies between experts. This is consistent with the extensive psychological literature on probability estimates in both lay and expert subjects (e.g.Doyle et al.,2011, 2014; Teigen and Brun, 1999; Bruine de Bruin et al., 2000; Budescu and Karelitz, 2004).


Next 3 years value ('A 1 in N chance')
10000000
1000000
100000
10000
1000
100
10
1
0
Extremely unlikely (1)
2
3
4
5
6
Extremely likely (7)
Qualitative likelihood of major eruption within 3 years



Next 30 years value ("A 1 in N chance")
100000
10000
1000
100
10
1
Extremely unlikely (1)
2
3
4
5
6
Extremely likely (7)
Qualitative likelihood of major eruption within 30 years

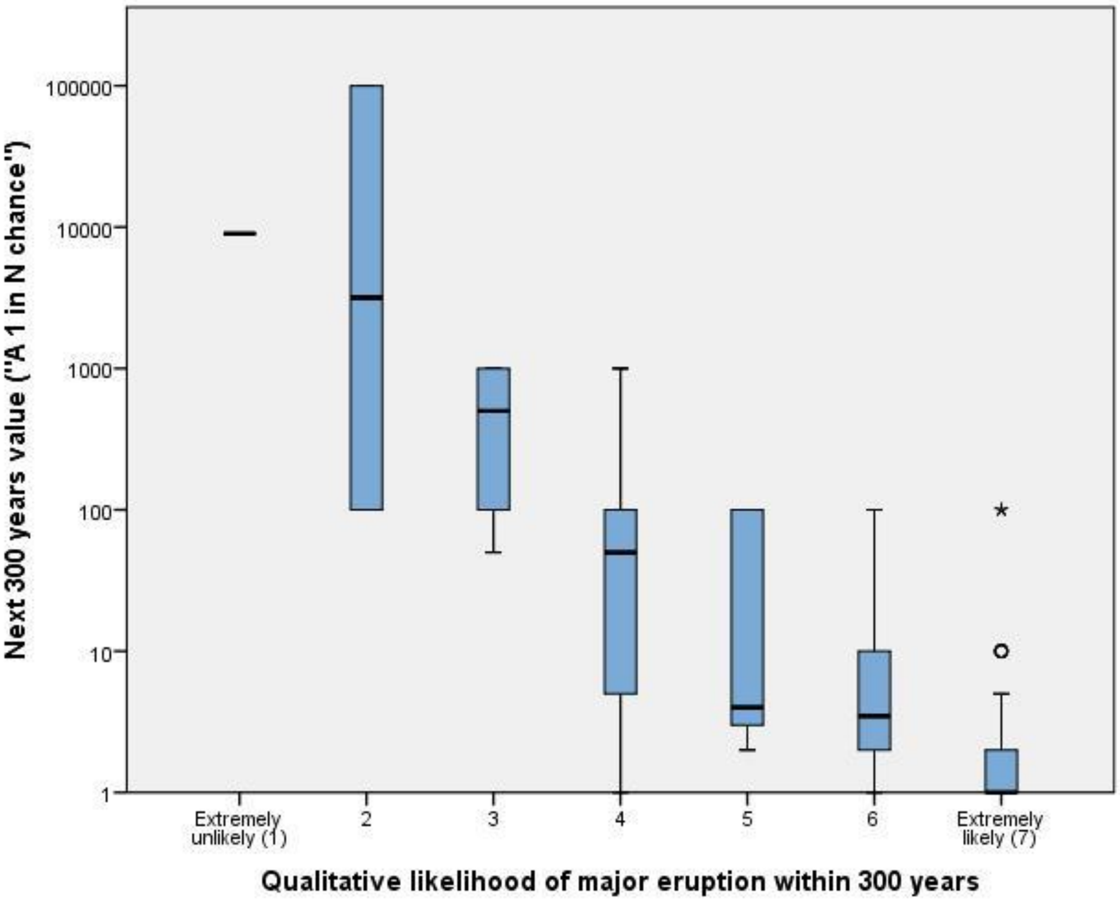


Figure 1. Range of values (1 in N) given for each time period, plotted against the position on the verbal scale selected by the respondents.

Use of the verbal scale in relation to chance estimates suggests (Figure 1) that there was considerable overlap between categories, particularly at the low probability end of the scale – where the overlap spans several orders of magnitude in the 3-year case (S3 and C3). However, over a 30-year timescale (S30 and C30), the overlap was more balanced with some overlap at the high-probability end of the scale. Finally, over the longest time period (S300 and C300), the variation was greatest at the low-probability end, but showed non-linear variation throughout. The majority of volcanoes were considered likely to erupt on a 300 year timescale (30 out of 70 volcanoes were considered to have a 1 in 1 chance of erupting; 55 were rated 6 or 7 on the qualitative scale).

4.3 Chance estimates

We asked scientists to express their probabilities in the form “1 chance in N”. Using this notation, ten scientists provided a higher value for the probability of eruption in 3 years than in 30 years (Check3-30), and 3 had provided a higher probability for an eruption in 300 years than in 30 years (Check30-300). The majority of scientists had successfully applied the notation, though as noted above 14 gave the same probability for all timescales, including 4 who gave unity for all timescales. There were no significant predictor variables.

A comparison between inconsistencies within the numerical likelihoods and those using the verbal scale showed that half of the inconsistencies were also replicated in the verbal scale between 3 and 30 years (there were no inconsistencies using the verbal scale between 30 and 300 years). This suggests that the chance notation itself was not the problem, but that estimating over specific time periods was. The respondents who made this error had also given high probabilities on both types of scale for 3 and 300 years, and then a much lower probability over 30 years. This result may imply that they assumed that the 30 year probability excluded the eruption they expect within 3 years – a question of phrasing. However, the verbal scale was in general used more consistently – and by all 95 respondents.

4.4 Use of scaling

The survey also allowed for analysis of the heuristics applied by scientists when scaling from a likelihood in 3 years (C3) to the other timescales. Scaling factors were calculated by dividing the longer timeframe values for “1 in N” with the shorter timeframes. The results suggest that almost four fifths of respondents used some form of scaling by an integer; 15 used a factor of 10 between 3 and 30 years, and 8 had used a factor of 10 between 30 and 300 years. Fourteen scientists gave the same probability for all timescales – six of these gave unity (i.e. certainty of an eruption). These data are summarised in Table 3.

	3 to 30 years	**30 to 300 years**
Linear scaling rule: factor of 2	10	13
Linear scaling rule: factor of 5	5	9
Linear scaling rule: factor of 10	14	9
Linear scaling rule: factor of 100	1	4
Other linear scaling rule	4	9
Non-linear but consistent rule (ie higher probability in longer timeframe)	14	13
Unity (i.e. same value for both)	14	14
Inconsistent	9	3
Total (N)	71	72

Table 3. Use of scaling between the chance estimates expressed as number of scientists using each rule. The reduction in N is due to three scientists not providing estimates for all timescales.

4.5 Timescale forecasts

In addition to the questions about likelihood over the three time periods, two further questions examined a different use of phrasing:

- “I believe that [volcano name] will certainly have a major eruption in the next A years” (T1)

- “I believe that there is at least a 50-50 chance of a major eruption at [volcano name] in the next B years” (T50)

Of the scientists who provided responses, 18 gave a number of years for the 50-50 chance that was higher than the “certainly” range. Nineteen had scaled up their 50-50 value by doubling it. Several scientists (9) gave the same response for each (T1 and T50).

4.6 Comparing chance and timescale estimates

In relation to the chance estimates, the distinctions were more marked. Thirty one respondents gave low probabilities for the likelihood of eruption in the next 300 years, but then gave much shorter timescales over which they were *certain* an eruption would occur. A further 7 gave a higher probability (unity) for the chance estimates and then a longer timescale in response to the “certainly” question (Consist_CT1).

Inconsistencies were slightly more marked for the “50-50” variable: thirty two respondents gave lower probabilities for chance estimates than for the 50-50 variable (e.g. probability of an eruption in 300 years= 1 in 10, but there is a 50-50 chance of an eruption within 20 years; Consist_CT50). A further 5 gave higher probabilities. Inconsistencies between chance estimates and the 50-50 value (Consist_CT50), and chance estimates and the “certainly” value (Consist_CT1), correlated(ρ=0.351, p<0.01).

4.7 Experts’ confidence

In total, fifty respondents exhibited either inconsistency between the chance estimates and timescale forecasts or inconsistency in their use of frequency notation, or both. Responses to the question “how confident are you in the estimates you have provided above” (M=4.44, SD=1.22) only predicted one of the inconsistencies – that between the 50-50 value and the likelihoods(i.e.CT50; ρ=-0.248, p<0.05). Otherwise, there was no relationship between how confident the experts were in their answers and how internally consistent the answers were.

Level of education did not produce any statistically significant results for any of the variables considered in this *Results* section.

5 DISCUSSION

The data presented above have a number of limitations. They are of necessity limited to volcanoes that are monitored – many volcanoes are not monitored at all, and of those that are, there is considerable variation in the methods used. The sample size is relatively small and may not be representative of all volcanologists – for example, the survey was only available in English, Spanish, French and Italian. The formulation of the questions may necessarily introduce some biases, though the results were interpreted with this in mind. However, there are several useful results that we draw out in this section. Initially, we discuss how our results might aid the framing of questions in risk assessments. We then discuss the implications of the study for understanding the cognitive processes involved in probability judgements, and suggest some further work that might clarify this. Finally, we assess the broader implications for expert judgements about volcanic risk.

5.1 Framing the questions

The use of the verbal likelihood scale correlated best with chance estimates over medium timescales, and least well over long timescales. The lower likelihood end of the scale was associated with greater range in numerical values. Low probabilities have been identified as more challenging to work with (Kahneman and Tversky, 1973). In addition, as Teigen and Brun (1999) argue, the use of a verbal scale does not necessarily correlate well with numerical probabilistic assessments, perhaps because inherent within the wording are particular implications. There may also be cultural and geographical variations in what is considered a low probability: the scientists represented a wide variety of contexts, but discrimination by volcano is not possible in a small dataset. Since volcanic eruptions at many volcanoes are low probability high impact events, this result suggests that discussion of uncertainties and terminology associated with probability is an important aspect of risk assessment.

The fact that the majority of respondents struggled in some way with consistency illustrates that phraseology and understanding the type of response required is critical in framing probability judgements. Psychological research demonstrates that the framing of questions is very important in the interpretation of probabilistic assessments (Gigerenzer, 1991; Tversky and Kahneman, 1986). In this case, expressing a probability as a "1 in N chance" over a specific time-period provides lower probabilities than expressing the result as a number of years in which the probability reaches certainty or fifty percent. Nevertheless, the use of frequency notation was relatively internally consistent: this method of expressing probabilities was generally understood. The timescale variables (T variables) were also relatively internally consistent. It is not possible to assess whether or not these values represent consistent underestimation of the probability – or whether the timescale variables represent overestimation. Had a single measure of probabilistic judgement been used – i.e. either chance estimates or timescale forecasts – these differences would not have been apparent. A possible explanation for the differences between the likelihood estimates and the timescale values is that scientists were more reluctant to give the answer "1 in 1" for the 300 years range, but were more confident when asked for the timeframe within which an eruption would certainly occur. It may also be the case that scientists found it harder to make estimates based on an externally provided timescale (3, 30 or 300 years): the fact that nine scientists did not provide these estimates but did provide estimates for T1 and T50 might support this.

It is also significant that more scientists were willing to provide timescale values than chance values. This suggests that they were more comfortable thinking about the number of years within which an event would certainly happen than about providing a chance estimate (“1 in N”) over a specified timeframe. Two of these were working on volcanoes that are thought relatively unlikely to erupt (values provided for “certainly within” were >100,000) – but the others were all working at timescales of 10 to 1000 years for this question. This suggests that some scientists generally found it slightly easier to imagine a timeframe than to imagine a chance of “1 in N”.

5.2 Scaling

Expert judgement for probabilistic assessments involves the production of individual probability distributions for particular potential events (e.g. Aspinall, 2006). In response to the survey, experts were asked to produce chance estimates for the likelihood of an eruption in 3, 30 and 300 years. In most cases, some form of internally consistent scaling was applied. This suggests that experts viewed the difference in probability over time as obeying a numerical law – though the nature of that law must have varied between experts. Our study could not distinguish between using different laws informed by differences in general understanding of volcanic processes or by volcano-specific differences (here differences might be defendable on scientific grounds). There were no obvious demographic factors.

The use of integers as scaling mechanisms may indicate a level of anchoring in the estimation of probabilities: once the probability for the first value has been estimated, the other two values follow (e.g. Tversky and Kahneman, 1974; Plous, 1989). Scientists clearly viewed the logarithmic timescale as suggesting some kind of regular relationship with probability. The survey results suggest that some scalings were based on $\log_{10}$, others used a linear scale and others a different exponent. This suggests that there might be some impact of the anchoring bias in the results. Further work, including interviews with scientists about their reasoning, would enable investigation of this process.

5.3 Expert probability judgement

The presence of these inconsistencies and biases suggests that volcanologists vary in their approach to probability judgements. We have no data concerning the level of experience that our sample of experts had in making such judgements operationally, so there is no information about the importance of experience in this paper. However, the psychological literature on probability judgement allows some interpretation of the results in terms of volcanological expertise and general probabilistic reasoning ability.

This paper has tested the ability – and willingness – of experts in volcanology to make probabilistic estimates in an informal context (without explanation of the process or any potential substantive outcome from the estimates, such as evacuations). Other studies of formal and informal settings have suggested that there is variation between contexts in terms of how well calibrated experts may be (e.g. review in O'Hagan et al., 2006), and also that expertise in a particular discipline is no guarantee of good calibration (Kahneman et al., 1982; Slovic, 2000) Probability estimation is a different cognitive process to volcanological research. It depends on a mathematical and philosophical understanding of probability theory, and on an awareness of the potential for value judgements to impact reasoning (e.g. Krinitzsky, 1993 – less of an issue, perhaps, in this paper, because the values were not being applied in risk assessment). The inconsistencies and heuristics revealed in this paper may therefore suggest that probabilistic methods in volcanology require a level of training (O'Hagan et al., 2006, make a number of constructive suggestions in this regard. See also Kynn, 2008).

6 CONCLUSIONS

The data suggest that experts, like laypeople, may be affected by a range of heuristics when asked to put probabilities on eruption likelihood over different timescales. Some of these produce biases (such as anchoring), while others introduce inconsistency. It is also evident that the estimation of lower probabilities is more challenging than estimates of events that are perceived as relatively likely.

We find that:

- Many scientists anchored their estimates to the initial estimate, and scaled up values from there using a range of different scaling factors;
- Scientists also interpreted timescale variables differently depending on the phraseology that was used, producing some inconsistencies in their results;
- Low probabilities presented particular challenges;
- Around 20% of scientists were unwilling to make numerical estimates of probability at all, due to high uncertainty;
- Scientists' confidence in their answers was generally independent of whether or not the answers were internally consistent.

These findings are consistent with previous studies (e.g. Tversky and Kahneman, 1986; Lichtenstein et al., 1981; Budescu et al., 2009), and suggest that probability estimation requires a level of training and calibration. They also demonstrate that there is high uncertainty both in volcano forecasting and in the way that scientists interpret questions, and that some scientists might be unwilling to participate in quantitative risk assessment through probability estimation. The variation in responses and consistency demonstrates the vulnerability of experts to the same inconsistencies as made by members of the public in estimating probabilities, though perhaps not to the same degree (e.g. Bolger and Wright, 1994; Slovic et al., 1981; Fischoff et al., 1982; Rowe and Wright, 2001). The phrasing and framing of probability statements can have a significant impact on the results. In volcanology, experts are frequently dealing with very uncertain and often low probabilities of high impact events. The data in this paper suggest that many of the surveyed scientists are not comfortable with providing numerical assessments of likelihood. They may be more comfortable with verbal assessments that the likelihood is increasing over background levels, for example – suggested by the unwillingness of some who used the qualitative scale to provide quantitative estimates. One approach to this might be the use of fuzzy sets (e.g. Dubois et al., 1993; Zadeh, 1982).

Acknowledgements

This work was carried out as part of the “VOLDIES” project (#228064), funded by the European Research Council. The authors are grateful to the survey respondents for their participation, and particularly to the World Organisation of Volcano Observatories, GNS Science in New Zealand and Montserrat Volcano Observatory for help in testing the survey.

References

Aspinall, W. P., G. Woo, B. Voight, and P. J. Baxter. 2003. Evidence-based volcanology:

Application to eruption crises. *Journal of Volcanology and Geothermal Research* 128:273-285.

Aspinall, W. P., S. C. Loughlin, F. V. Michael, A. D. Miller, G. E. Norton, K. C. Rowley, R. S. J. Sparks, and S. R. Young. 2002. The Montserrat Volcano Observatory: its evolution, organization, role and activities. *Geological Society, London, Memoirs* 21 (1):71-91.

Aspinall, W.P.. 2006. Structured elicitation of expert judgement for probabilistic hazard and risk assessment in volcanic eruptions. In *Statistics in Volcanology*, ed. S. C. H.M. Mader, C. Connor, L. Connor, 15-30. London: Geological Society of London.

Bolger, F., and G. Wright. 1994. Assessing the quality of expert judgment: Issues and analysis. *Decision Support Systems* 11 (1):1-24.

Bruine De Bruin, W., B. Fischhoff, S.G. Millstein, B.L. Halpern-Felsher. 2000. Verbal and numerical expressions of probability: “It's a fifty-fifty chance.” *Organ. Behav. Hum. Decis. Process*., 81 (1) (2000), pp. 115–131

Brun, W., K.H. Teigen 1988. Verbal probabilities: ambiguous, context-dependent, or both? *Organ. Behav. Hum. Decis. Process.,* 41 pp. 390–404

Budescu, D. V., Broomell, S., & Por, H. H. (2009). Improving communication of uncertainty in the reports of the Intergovernmental Panel on Climate Change. *Psychological Science*, 20(3), 299-308.

Budescu, D.V., T.M. Karelitz, T.S. Wallsten Predicting the directionality of probability words from their membership functions *J. Behav. Decis. Mak*., 16 (3) (2003), pp. 159–180.

Burgman, M., Carr, A., Godden, L., Gregory, R., McBride, M., Flander, L., & Maguire, L. 2011. Redefining expertise and improving ecological judgment. *Conservation Letters*, *4*(2), 81-87.

Burgman, M. A., McBride, M., Ashton, R., Speirs-Bridge, A., Flander, L., Wintle, B., Fidler, F., Rumpff, L. & Twardy, C. (2011). Expert status and performance. *PLoS One*, *6*(7), e22998.

Cashman, K. V., & Sparks, R. S. J. (2013). How volcanoes work: A 25 year perspective. Geological Society of America Bulletin, 125(5-6), 664-690.

Cooke, R.M. 1991: *Experts in uncertainty: Opinion and subjective probability in science* Oxford: Oxford University Press.

Dawid, A. P. 1982. The Well-Calibrated Bayesian. *Journal of the American Statistical Association* 77 (379):605-610.

Donovan, A., C. Oppenheimer, and M. Bravo. 2012. The use of belief-based probabilistic methods in volcanology: Scientists' views and implications for risk assessments. *Journal of Volcanology and Geothermal Research*.

Donovan, A., Eiser, J. R., & Sparks, R. S. J. (2014). Scientists' views about lay perceptions of volcanic hazard and risk. *Journal of Applied Volcanology*, 3(1), 1-14.

Donovan, A., & Oppenheimer, C. (2015). At the mercy of the mountain? Field stations and the culture of volcanology. *Environment and Planning A*, 47(1), 156-171.

Doyle, E. E., Johnston, D. M., McClure, J., & Paton, D. (2011). The communication of uncertain scientific advice during natural hazard events. *New Zealand Journal of Psychology*, 40(4), 39-50.

Doyle, E. E., McClure, J., Johnston, D. M., & Paton, D. (2014). Communicating Likelihoods and Probabilities in Forecasts of Volcanic Eruptions. *Journal of Volcanology and Geothermal Research.*

Dubois, D. J., Prade, H., & Yager, R. R. (Eds.). (1993). *Readings in fuzzy sets for intelligent systems*. Morgan Kaufmann.

Eiser, J. R., & White, C. J. (1974). Evaluative consistency and social judgment. *Journal of*

Personality and Social Psychology, 30(3), 349.

Eiser, J. R., & Hoepfner, F. (1991). Accidents, disease, and the greenhouse effect: effects of response categories on estimates of risk. *Basic and Applied Social Psychology*, 12(2), 195-210.

Ferrell, W. R. 1994. Calibration of sensory and cognitive judgments: A single model for both. *Scandinavian Journal of Psychology*, *35*(4), 297-314.

Fischhoff, B. 1991. Value elicitation: is there anything in there?. *American Psychologist*, *46*(8), 835.

Fischhoff, B., P. Slovic, and S. Lichtenstein. 1982. Lay foibles and expert fables in judgments about risk. *The American Statistician* 36 (3b):240-255.

Gigerenzer, G. 1991. How to Make Cognitive Illusions Disappear: Beyond 'Heuristics and Biases' *European Review of Social Psychology* 2 (1):83-115.

Gigerenzer, G. 1994. Why the distinction between single-event probabilities and frequencies is relevant for psychology (and vice versa). Pages 129-162 *in* G. Wright and P. Ayton, editors. *Subjective probability.* Wiley, New York, New York, USA.

Gigerenzer, G. 2008. Why Heuristics Work. *Perspectives on Psychological Science* 3 (1):20.

Gigerenzer, G., R. Hertwig, E. V. D. Broek, B. Fasolo, and K. V. Katsikopoulos. 2005. 'A 30% chance of rain tomorrow': How does the public understand probabilistic weather forecasts? *Risk Analysis* 25 (3):623-629.

Gigerenzer, G., W. Gaissmaier, E. Kurz-milcke, L.M. Schwartz, S. Woloshin. 2007. Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest* 8 (2):53-96.

Goldstein, D. G., & Rothschild, D. (2014). Lay understanding of probability distributions. *Judgment and Decision Making*, *9*(1), 1-14.

Hacking, I. 2001. *An introduction to probability and inductive logic*. Cambridge: Cambridge University Press.

Hincks, T.K., Komorowski, J-C., Sparks, R.S.J. and Aspinall, W.P. Retrospective analysis of uncertain eruption precursors at La Soufrière volcano, Guadeloupe, 1975-77: volcanic hazard assessment using a Bayesian Belief Network approach. *Journal of Applied Volcanology* 3:3, 2014

Hoffrage, U., S. Lindsey, R. Hertwig, and G. Gigerenzer. 2000. Communicating statistical information. *Science* 290 (5500):2261-2262.

Kahneman, D., and A. Tversky. 1972. Subjective probability: A judgment of representativeness. *Cognitive Psychology* 3 (3):430-454.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological review*, 80(4), 237.

Kahneman, D., & Tversky, A. (1982). On the study of statistical intuitions. *Cognition*, *11*(2), 123-141.

Kahneman, D., & Tversky, A. (1996). On the reality of cognitive illusions. *Psychological Review*, Vol 103(3), Jul 1996, 582-591

Kahneman, D., P. Slovic, and A. Tversky. 1982. *Judgment under uncertainty: Heuristics and biases*: Cambridge University Press.

Karelitz, T. M., & Budescu, D. V. (2004). You say" probable" and I say" likely": improving interpersonal communication with verbal probability phrases. Journal of Experimental Psychology: Applied, 10(1), 25.

Krinitzsky, E. L. (1993). Earthquake probability in engineering—Part 1: The use and misuse of expert opinion. The Third Richard H. Jahns Distinguished Lecture in engineering geology. *Engineering Geology*, *33*(4), 257-288.

Kynn, M. 2008. The ‘heuristics and biases’ bias in expert elicitation. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 171 (1):239-264.

Lichtenstein, S., and B. Fischhoff. 1980. Training for calibration. *Organizational Behavior and Human Performance* 26 (2):149-171.

Lichtenstein, S., B. Fischhoff, and L. D. Phillips. 1981. Calibration of probabilities: The state of the art to 1980: DTIC Document.

Lindsay, J., W. Marzocchi, G. Jolly, R. Constantinescu, J. Selva, and L. Sandri. 2010. Towards real-time eruption forecasting in the Auckland Volcanic Field: application of BET_EF during the New Zealand National Disaster Exercise ‘Ruaumoko’. *Bulletin of Volcanology* 72 (2):185-204.

MacGillivray, B. H. 2014. Heuristics Structure and Pervade Formal Risk Assessment. *Risk Analysis*, 34(4), 771-787.

Martin, T. G., Burgman, M. A., Fidler, F., Kuhnert, P. M., LOW‐CHOY, S. A. M. A. N. T. H. A., McBride, M., & Mengersen, K. 2012. Eliciting expert knowledge in conservation science. *Conservation Biology*, *26*(1), 29-38.

Marzocchi, W., L. Sandri, J. Selva. 2008. BET_EF: A probabilistic tool for long- and short-term eruption forecasting. *Bulletin of Volcanology* 70 (5):623-632.

Marzocchi, W., Newhall, C. and Woo, G. 2012: The scientific management of volcanic crises. *Journal of Volcanology and Geothermal Research* 247-248, 181-189.

Marzocchi, W., and G. Woo. 2009. Principles of volcanic risk metrics: Theory and the case study of Mount Vesuvius and Campi Flegrei, Italy. *J. Geophys. Res.* 114.

Marzocchi, W., L. Sandri, P. Gasparini, C. Newhall, and E. Boschi. 2004. Quantifying probabilities of volcanic events: The example of volcanic hazard at Mount Vesuvius. *Journal of Geophysical Research* 109.

Marzocchi, W., L. Sandri, J. Selva 2008: BET_EF: A probabilistic tool for long- and short-term eruption forecasting. *Bulletin of Volcanology* 70, 623-632.

Murphy, A. H., & Winkler, R. L. 1984. Probability forecasting in meteorology. *Journal of the American Statistical Association*, *79*(387), 489-500.

Neri, A., W. P. Aspinall, R. Cioni, A. Bertagnini, P. J. Baxter, G. Zuccaro, D. Andronico, S. Barsotti, P. D. Cole, T. Esposti Ongaro, T. K. Hincks, G. Macedonio, P. Papale, M. Rosi, R. Santacroce, and G. Woo. 2008. Developing an Event Tree for probabilistic hazard and risk assessment at Vesuvius. *Journal of Volcanology and Geothermal Research* 178 (3):397.

Newhall, C., and R. P. Hoblitt. 2002. Constructing event trees for volcanic crises. *Bulletin of Volcanology* 64:3-20.

O'Hagan, A., C. E. Buck, A. Daneshkhah, J.R. Eiser, and P. Garthwaite. 2006. *Uncertain judgements: Eliciting experts' probabilities*. London: Wiley.

Parducci, A. (1965). Category judgment: a range-frequency model. *Psychological review*, 72(6), 407.

Plous, S. 1989. Thinking the Unthinkable: The Effects of Anchoring on Likelihood Estimates of Nuclear War. *Journal of Applied Social Psychology* 19 (1):67-91.

Risbey, J. S., & Kandlikar, M. 2007. Expressions of likelihood and confidence in the IPCC uncertainty assessment process*. Climatic Change*, 85(1-2), 19-31.

Rowe, G., & Wright, G. 2001. Differences in expert and lay judgments of risk: myth or reality? Risk Analysis, 21(2), 341-356.

Slovic, P., B. Fischhoff, S. Lichtenstein, and F. J. C. Roe. 1981. Perceived Risk: Psychological Factors and Social Implications [and Discussion]. *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences* 376 (1764):17-34.

Slovic, P. E. 2000. The perception of risk. Earthscan Publications.

Smith, J. F., & Kida, T. 1991. Heuristics and biases: Expertise and task realism in auditing. *Psychological Bulletin*, *109*(3), 472.

Sparks, R. S. J. (2003). Forecasting volcanic eruptions. Earth and Planetary Science Letters, 210(1), 1

Sparks R.S.J., Aspinall, W.P., Crosweller, H.S. and Hincks, T.K. Risk and Uncertainty assessment of volcanic hazards 2013. In "*Risk and Uncertainty Assessment of Natural Hazards*" eds. Rougier, J., Sparks, R.S.J. and Hill, L. Cambridge University Press, Chapter 11, 364-397, 2013.

Teigen, K. H., and W. Brun. 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 expressions of uncertainty and probability. *Thinking: Psychological perspectives on reasoning, judgment and decision making*, 125-145.

Tilling, R. I. (2008). The critical role of volcano monitoring in risk reduction. *Advances in Geosciences,* 14(14), 3-11.

Tversky, A., and D. Kahneman. 1973. Availability: A heuristic for judging frequency and probability. *Cognitive Psychology* 5 (2):207-232.

Tversky, A., and D. Kahneman 1974. Judgment under Uncertainty: Heuristics and Biases. *Science* 185 (4157):1124-1131.

Tversky, A., & Kahneman, D. 1983. Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological review*, *90*(4), 293.

Tversky, A., & Kahneman, D. 1986. Rational choice and the framing of decisions. *Journal of business*, S251-S278.

Vranas, P. 2000. Gigerenzer's normative critique of Kahneman and Tversky. *Cognition* 76 (3):179-193.

Wynne, B. (1996). A reflexive view of the expert-lay knowledge divide. *Risk, environment and modernity: towards a new ecology*. Sage, London, 44-83.

Zadeh, L. A. (1982). Fuzzy probabilities and their role in decision analysis. In Proc. of the IFAC Symp. on Theory and Application of Digital Control (pp. 15-23).