Deliverable D5.4

Field evaluation of OEF communications

Deliverable information	
Work package	WP5: Society: Data Gathering and Information Sharing with the Public and Policy-makers
Lead	University of Cambridge
Authors	Sarah Dryhurst, Irina Dallo, Giulia Luoni, Michele Marti, Alexandra L.J. Freeman
Reviewers	Rémy Bossu, EMSC
Approval	Management Board
Status	Final
Dissemination level	Public
Will the data supporting this document be made open access? (Y/N)	Y
If No Open Access, provide reasons	
Delivery deadline	30th August 2022
Submission date	29th August 2022
Intranet path	

29/08/2022

This project has received funding from the European Union's Horizon 2020 research and innovation

Table of Contents

1. Summary

In order to provide advice to countries that already, or will soon, have Operational Earthquake Forecasts (OEFs), we have been investigating how best to communicate such forecast information. OEFs are already publicly available in New Zealand and have occasionally been made public in Iceland, whilst in the US, operational aftershock forecasts are often published (Dryhurst *et al.*, 2020). In Italy and Japan OEFs are not publicly available, although there are moves to make them so, and they are currently under development in Switzerland.

In this study, we carried out three experiments across three countries with different seismic hazard levels (California in the US, Italy and Switzerland), using three different languages (English, German and Italian). In each country we surveyed at least 2732 people, giving us adequate power to test four experimental arms against each other within each of these three experiments. The aim was to evaluate the effects of different ways of communicating probabilities in operational earthquake forecasting.

The three experiments revealed:

- That **people in all three countries gave very similar answers** across all questions, except that those in Switzerland (the country with the lowest seismic hazard of the three) had a lower perception of the forecast risk and were correspondingly less likely to say that they would take any action in preparation for an earthquake as a result of an operational earthquake forecast. Further analysis of the data will reveal whether this is due to a lower proportion of people in Switzerland having experienced an earthquake.

- That **maps attempting to represent operational earthquake forecast probabilities as different coloured isoline compartments appear either not to be used by, or possibly even mislead, public audiences**. Although they give 'at a glance' information for a policy-maker with a large geographical area to consider, for an individual interested in only their own area, a simple geographic map illustrating the area over which the forecast is valid is probably all that is needed.

- That presenting the forecast probability with a phrase such as 'out of 100,000 towns with exactly this chance of suffering a damaging earthquake we would expect…' leads to a significantly higher perception of the risk than presenting it as an absolute percentage chance alone, at realistically low probabilities. *Combining* percentage chance with this kind of 'expected frequency' format was rated as easier to understand than the expected frequency alone (although harder than absolute percentage alone). **Presenting probabilities as both percentage and expected frequency together might be considered best practice as the expected frequency format helps explain the absolute percentage and heightens discrimination between very low percentages.**

- Presenting the baseline risk for the forecast location alongside the current forecast as a percentage appeared to have different effects in different countries. It tended to increase the risk perception compared to the current absolute forecast risk alone for participants in Switzerland, however for those in Italy it reduced their perception of the risk at higher hazard levels. **Because of this unexplained, variable effect we**

advise against presenting the local baseline hazard level alone as a way of trying to give context to the forecast numbers.

- Presenting a relative risk (calculated from the comparison between the current forecast and the baseline risk for the area) alongside the absolute forecast percentage heightened people's perception of risks above baseline, but reduced discrimination between these above-baseline probabilities (in Italy at least). This format was also ranked poorly by participants. **We therefore do not recommend using a relative risk combined with the absolute forecast risk as a way of giving greater context.**

- Giving participants a risk ladder depicting comparator risks for context was appreciated by some but not all participants. Out of the two types of risk ladders we evaluated, the one that presented the seismic hazard in other, familiar, cities increased perception of the risk the most, and was generally preferred by participants in their rankings of the formats. However, such risk ladders were deemed quite complex to understand. **We recommend that showing the current local forecast probability positioned along a risk ladder with the baseline probability in other, familiar, cities as comparators is offered only as secondary information.**

The combination most likely to be helpful to a public audience, then, is:

- A geographical map showing the area of the forecast.
- The current, absolute percentage chance of an earthquake in the defined forecast period together with that chance expressed in an expected frequency format ('Out of 100,000 towns with exactly this chance, we would expect…').
- Some audience members would appreciate an option to put the forecast chance in context, leading to a risk ladder illustrating the current local risk alongside the background seismic hazard in a range of familiar locations.

See Figure 1 for a representation of this combination.

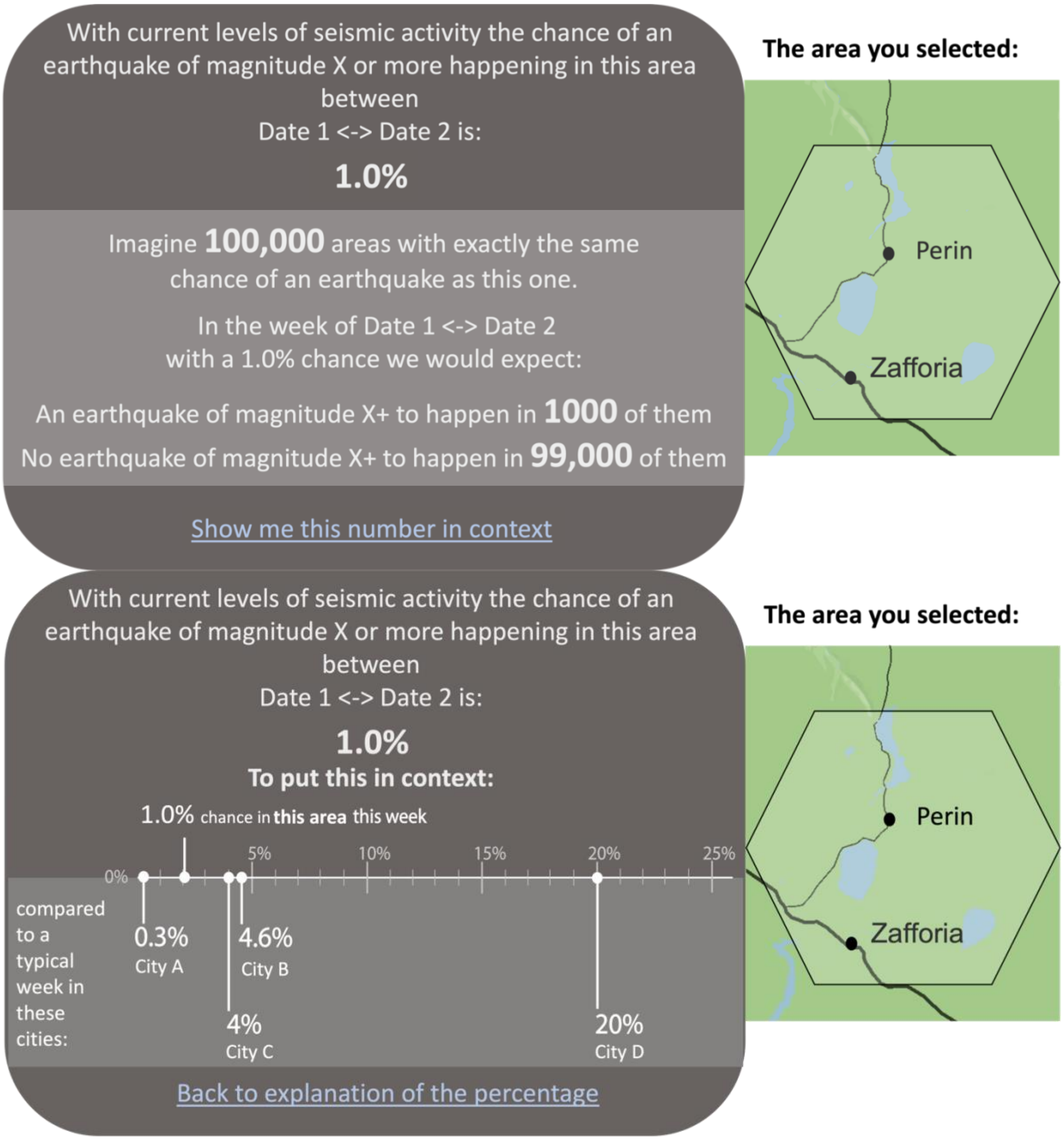


Figure 1: Suggested best practice for main page (top) and click-through secondary information (bottom) to communicate OEF probabilities to a public audience (N.B. exact wording around communicating the area of the forecast, the size of the earthquake being forecasted, and optional navigation between primary information and further context not tested in this study).

2. What aspects of the communication are being evaluated?

Whilst developing potential dashboards to communicate operational earthquake forecasts, we came across a few key issues:

- **Information overload.** We want to be able to give people the minimum amount of information necessary to correctly understand and interpret the forecast.

- **Clarity of purpose.** We need to make the purpose of each piece of information included clear, and to set people's expectations of why they are being given information.

- **Probabilistic information.** We need to help people understand that a forecast is inherently uncertain, but also to help them understand what the information means and why it can still be helpful, even when uncertain.

- **Giving context.** We need to help people understand what the probability actually means for them.

To address these issues, we designed three quantitative experiments to ascertain the best ways of communicating a key OEF forecast probability in a way that was clear, concise and helpful in terms of people understanding that the forecast was probabilistic, and what it meant in terms of risk to them.

In these experiments we wanted to concentrate on the probabilistic information, not on how best to communicate the magnitude or intensity relating to the forecast, so we simplified the wording to simply 'a damaging earthquake' as we thought that wording would be the most likely to be interpreted with a similar level of concern across all participants (whilst 'a felt earthquake' for example, would likely be much more worrying to some than to others, depending on their previous experience and the robustness of their surrounding architecture).

The three experiments were as follows:

1) The first experiment was designed to address **how to present the forecast probability** itself: as a percentage or as an expected frequency, or both, and whether the addition of a graphical aid helped people comprehend the size of that number.

2) The second experiment was designed to address **how best to add context** to the forecast probability to help people interpret it: by adding the typical, baseline probability for the forecast location as a comparison; by adding a relative risk (based on the difference between the forecast probability and the typical baseline probability for the area); by adding a risk ladder with other risk comparators; or by adding a risk ladder with other cities' seismic hazards as comparators.

3) The third experiment was designed to address the issue of including a **map** alongside the forecast. Maps prove very popular with different audiences for forecast information – they are familiar and people expect one to accompany a 'forecast' much as they would with a weather forecast. We had concerns about how seeing the forecast hazard levels of surrounding areas might affect people's responses to the forecast hazard level at their own location, so in this third experiment we compared different ways of communicating using a map.

In all experiments, the main thing we wanted to know was **how the forecast format people saw affected the way they perceived the probability presented**. We therefore gave each participant several hazard levels presented in the same format and asked them 'How would you classify that risk in your mind?', asking them to indicate their answer on a slider from 'very low risk' to 'very high risk'. This allowed us to compare people's answers at different hazard levels, as well as across different formats.

We also wanted to know **how easy or difficult people felt the information was to understand**, and asked them this once per format, per experiment, giving us a rating for each format.

We also collected information on **how worried they would be if shown that information**, to see whether some formats made people more worried than others.

For the experiment on the effects of different contextual information types, which is closest to how we imagine the final presentation in a forecast dashboard looking, we also asked **what actions they would take as a response if they saw the forecast**, which we hoped would act as an additional measure of their level of worry and help us to calibrate what might be an 'appropriate' level of perception of risk and worry in the real world.

It is very important to note that we are not trying to find the format that makes people perceive the risks as the highest, or the lowest, but ideally the format that helps them discriminate: that makes the lowest risks seem lowest and the highest risks seem highest. We also want to use the reported actions that people said they would take to ensure that they are 'appropriate' for those lowest and highest hazard levels in particular.

All these experiments were carried out in three different areas with varying background hazard levels: Italy, California and Switzerland. We collected information about exactly where people lived within these locations, and their prior experience with earthquakes as well as information about their beliefs and knowledge about earthquakes, and demographic information including measuring their numeracy, optimism/pessimism and fatalism beliefs, any of which might affect their perceptions and answers to our questions. We also ensured that our samples were large and representative of the relevant country in terms of both age and sex breakdown. For full details of the demographics, see Table A1.

2.1. Methods that apply to all experiments

All three experiments were contained within a single survey, one for each country in turn. Participants were recruited through ISO-accredited online survey providers (Dynata for Italy and California, and Respondi for Switzerland), and completed the survey in Qualtrics. The survey took 20-25 minutes to complete and participants were paid a set fee depending on the country they were in. All participants were 18+ and were recruited by quotas proportional to the age and sex demographics of the country they were in, according to the most recently available census (2020 in US and Italy, 2021 in Switzerland).

A total of 2732 participants were recruited per country based on a power calculation in GPower, based on a linear model with a categorical predictor having a small effect (d = 0.2), using an overly-conservative Bonferroni correction accounting for 6 multiple comparisons – one for each combination of the 4 arms in the experiment (alpha = 0.05/6 ~ 0.008) and 85% power (n = 683 per arm, n = 2732 in total). In analyses, we use the Benjamini-Hochberg procedure to keep the false discovery rate at 5%.

The survey included two attention checks. In California this included one simple question ("*To check that you are paying attention, please select 'Disagree' below.*"), and one more difficult question (*"How concerned are you about covid-19? To show us that you're still paying attention, please select 'somewhat concerned' below."*). Due to the difficulty of this second one, in Switzerland and Italy it was replaced with *"How concerned would we be if you weren't paying attention? To show us that you're still paying attention, please select 'somewhat concerned' below."* Given we wanted to compare between countries for the purposes of this report, all analyses presented here exclude participants who failed the first attention check question (which was consistent between countries), giving us n=2290 for California, n=2383 for Switzerland, n=2326 for Italy.

3. Experiment 1: Evaluating different formats for communicating the likelihood of an earthquake

3.1. Introduction

When communicating the probability of an event happening, the format that the probability is presented in can change the audiences' perception of how likely the event is to occur. For example, probabilities represented as percentages (e.g. 10%) seem less likely than the same probabilities represented as a natural or expected frequency (how many times it might be expected to happen out of a set number of opportunities, e.g. 10 out of 100) (Peters, Hart and Fraenkel, 2011; Freeman *et al.*, 2021). This effect has sometimes been shown to vary depending on the numeracy of the audience (Peters, Hart and Fraenkel, 2011).

The probability of a seismic event happening, in a defined area, during a defined period of time – such as is communicated as part of an operational earthquake forecast – ranges over several orders of magnitude. During an earthquake swarm or aftershock sequence it can near 100%, whilst during a quiescent period in a low hazard area it can be lower than 0.001% (1 in 100,000). Most of the time then, an OEF will be trying to communicate very low probabilities. This increases the difficulty of the choice of format to present the numbers in, since percentages with a lot of decimal places may well be harder to conceive of in terms of size than their equivalent expected frequency format.

People tend to find all probabilities a difficult concept as they involve inherent uncertainty and are actually stimulating the imagination of different possible future scenarios. Expected frequencies make that much more explicit than a percentage chance – spelling out the number of different alternative outcomes (e.g. 10 times in which the event occurs, out of 100 times overall). This might explain their popularity with audiences, and in fact the usual way in which expected frequencies are used almost removes the uncertain future component of the probability, and collapses it to a more certain outcome: 'out of 100 people facing this event, we'd expect it to happen to 10 of them'. The ability to bring the event to more vivid imagination and possibly the decrease of the feeling of uncertainty may explain why events expressed as expected frequencies are often perceived as more likely to happen than those expressed as a probability (Siegrist, 1997; Slovic, Monahan and MacGregor, 2000; Keller, Siegrist and Gutscher, 2006).

Low probabilities are even more difficult to communicate (Camerer and Kunreuther, 1989; Halpern, Blackman and Salzman, 1989; Lipkus, 2007), and the likelihoods of rare or global natural hazards (such as a damaging earthquake, volcano, meteor strike or global warming) are also more difficult to express as expected frequencies. These require imagining many possible futures involving the same geographical location, or imagining many different similar geographical locations (e.g. multiple earth-like planets) with the same hazard level, which may be more difficult, conceptually, than the more common task of imagining many different people facing the same hazard. With a local earthquake forecast, however, the geographical area that the forecast is valid over is small enough that imagining many similar geographical locations (in this case, for example, 'towns') is perhaps not so conceptually difficult.

Graphical representations can help people understand numbers, and a stacked bar is particularly recommended for helping people understand a part-to-whole relationship such as a percentage (Lipkus, 2007). During the development phase of the OEF communications we experimented with both pie and bar designs, and the members of the public we consulted generally preferred the bar design, so we tested the effects of that in this experiment. It is difficult to use a single bar to represent a wide range of values, including very small percentages, but whilst mindful of that difficulty we wanted to test its effects quantitatively. In later experiments in this study we test other graphical and contextual aids.

In this experiment, therefore, we are trialing several potential ways of representing likelihoods: in percentages or expected frequencies and with or without a graphical aid to help represent the number, or a combination of percentages and expected frequencies together. We are using likelihoods that span a realistic range of hazard levels (from 0.001% to 44%), in order to assess how people perceive the risk, and how much worry that stimulates in them, at each hazard level.

3.2. Methods

Participants were randomised into one of four experimental arms, each showing a different format. Each participant saw five stimuli in a randomised order within their respective arm, with each stimulus showing the arm's respective format but at a different hazard level: 0.001%, 0.02%, 1%, 22%, 44%. These hazard levels were chosen to represent a realistically broad range of probabilities that might be relevant for OEF. The 22% and 44% levels were chosen to overlap with those hazard levels used in Experiment 3, to allow informal comparisons between the effects of the formats used in the two different experiments.

The four arms' formats were chosen from a set of five possibilities, with different comparisons made in different countries, to allow us to compare all five different formats whilst only using four arms per country (as our experimental power permitted us to do). See Figure 2 for the graphics used in each country.

Arm 1 – 'Future frequency' format
Expected frequencies in 'imagine possible futures' format *('Imagine 100,000 possible future ways in which the week of 6th July <-> 13th July could turn out in your town')*.

Arm 2 – 'Geographical frequency' format
Expected frequencies in 'geographical' format *('Imagine 100,000 towns with exactly the same chance of an earthquake as your town in the week 6th July <-> 13th July')*.

Arm 3 – Percentage format

Arm 4a – Percentage/geographical frequency format
Percentage combined with 'geographical' frequency format (used in Italy and Switzerland).

Arm 4b – Percentage format with bar graphic
Percentage format accompanied by a graphical representation in the form of a bar (used in California).

We collected data on four main dependent variables:

1) **Risk perception**:
- *Imagine that you visited a website that showed you this graphic indicating the chance of a damaging earthquake happening in your town within the next 7 days. How would you classify that risk in your mind?*
    - Move the slider below to indicate it: 'Very low risk' – 'Very high risk')

2) **Subjective comprehension** (measured via two questions, only once per arm, at the 22% hazard level):
- *How clear and easy did you find the graphic to understand?*
    - Move the slider below to indicate it: 'Not at all clear' – 'Very clear'

- *How much effort did you feel you had to put into understanding the graphic?*
    - Move the slider below to indicate it: 'A lot of effort' - 'Very little effort'

3) **Worry** (measured only once per arm, at the 22% hazard level):
- *How worried would you be about the chance of a damaging earthquake happening in your town during the forecast period (6th-13th July) if you were shown this graphic?*
    - Move the slider below to indicate it: 'Not at all worried' – 'Very worried'

4) **Surprise** (understanding of uncertainty) (measured only once per arm, at the 22% hazard level and only in California):
- *How surprised would you be if a damaging earthquake happened in your town during the forecast period (6th-13th July) if you had seen this graphic?*
    - Move the slider below to indicate it: 'Not at all surprised' – 'Very surprised'

We additionally collected potential covariates including education, objective numeracy, optimism/pessimism, fatalism and earthquake experience.

We pre-registered four formal hypotheses at https://osf.io/5kf9r/:

1) Participants in all arms seeing five graphics representing different hazard levels will perceive the risks to be higher in the graphics showing the higher hazard levels.

2) Participants seeing hazards represented as expected frequencies will perceive the risks as higher than those seeing the hazards represented as percentages.

3) Participants seeing hazards represented as expected frequencies will rate their worry as higher than those seeing the hazards represented as percentages. This will only be tested at the 22% hazard level where the worry question is asked. (Preregistered for Switzerland and Italy only).

4) Participants seeing hazards represented as a future frequency (imagining possible futures) will rate the information subjectively harder to understand than those seeing hazards represented as a geographical frequency (imagining identical towns), and participants seeing either of these will rate the information as subjectively harder to understand than those who saw the hazards represented as a percentage (with or without a graphical representation) (note that this will only be tested for the one hazard level where the subjective comprehension questions are asked i.e. 22%).

Figure 2: graphics used in Experiment 1 in a) California (the percentage/geographical frequency format wasn't used in the experiment but is included for the English translation), b) Switzerland and c) Italy

3.3. Results

Risk perception

Figure 3 shows how risky participants rated each hazard level in each format as feeling (hazard levels are within-subjects, formats between-subjects).

It is clear from the data that our first two hypotheses are supported: participants in all arms seeing the five graphics representing different hazard levels (within subjects) perceived the risks to be higher in the graphics showing the higher hazard levels. In turn, participants seeing the hazards represented as expected frequencies (both the ‘geographical’ and ‘future’ frequency formats) perceived the risks as higher than those seeing the hazards represented as percentages (between subjects). The exception to this confirmation of our second hypothesis was at the 44% hazard level, where the differences between the expected frequency formats and the percentage format become much less obvious. Although the ‘future’ frequency format does appear to be perceived as slightly less risky than the percentage format at the 44% hazard level in Switzerland and Italy.

Comparing the two frequency formats (‘future’ and ‘geographical’) in more detail, the ‘future’ frequency format appears to be perceived as slightly less risky than the ‘geographical’ frequency format – possibly because it is less easy to imagine psychologically. It should be noted that these differences are not significant for California or for the lower hazard levels (0.001%, 0.02% and 1%) in Italy however.

The data from California, which included an arm showing a bar graphic alongside a percentage, suggests that the bar graphic made no difference to the perception of the forecast probability when that probability was very small. This was perhaps to be expected, since the bar graphic was not able to illustrate a difference between percentages smaller than 1%. At hazard levels of 1% and above, participants seeing the percentage with a bar graphic were able to recognise the difference between different hazard levels, and in these cases, rated the higher hazard levels as less risky than those seeing the percentage without a bar graphic.

In Italy and Switzerland, this bar graphic arm was replaced by one that showed a format that combined the ‘geographic’ frequency format with the percentage format. Due to an error in the survey coding, participants in Italy were shown the wrong graphic for the 22% hazard level in this arm, and thus were excluded from the present analyses. Regardless, the available results in both countries clearly show that the arm combining the ‘geographic’ frequency and percentage formats gives a higher perception of the risk than that which shows the percentage format alone, except in Italy at the 44% hazard level, where there is no significant difference (as discussed above). This combined ‘geographical’ frequency and percentage arm also appears to give a lower perception of the risk than for the ‘geographical’ frequency format alone for hazard levels at 1% and below.

Interestingly, across all three countries, at the highest hazard level (44%) all the formats are much more similar in their risk perception.

It's also interesting to note that in Switzerland, the country with the lowest seismic hazard of the three, all risk ratings are generally lower than in California or Italy (which are both high hazard areas).

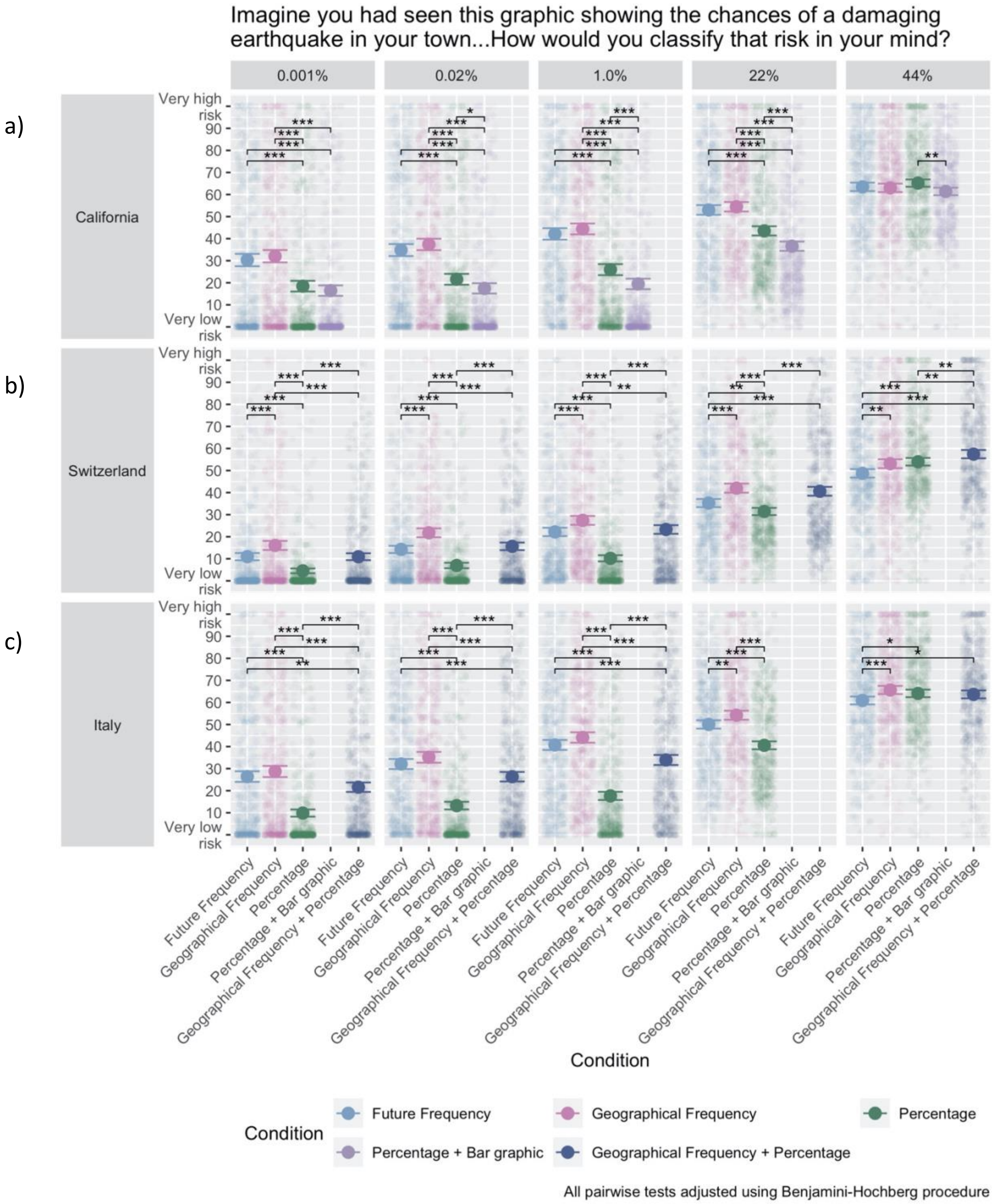


Figure 3: Participants' risk perceptions of the different formats and hazard levels in a) California (n=2290), b) Switzerland (n=2383) and c) Italy (n=2326). Note – not all formats were tested in all countries, hence the blank columns. Individual data points are plotted as jittered points, with means shown as dots with confidence intervals (+/- 95% CIs). Asterisks denote significant pairwise differences between formats, adjusted using the Benjamini-Hochberg procedure (p <.05; ** p <.01, *** p <.001 – all are considered 'significant' in the results analysis).*

Subjective comprehension

Figure 4 shows participants' subjective feelings about how easy each format was to understand, and relatedly how much effort they felt they had to put into understanding these formats.

These results partly support our fourth hypothesis: regarding subjective perceptions of 'future' vs 'geographical' frequency formats, in line with our hypothesis, participants seeing hazards represented as a 'future' frequency (imagining possible futures) rated the information as subjectively harder to understand than those seeing hazards represented as a 'geographical' frequency (imagining identical towns) in California and Italy, although this was not the case in Switzerland. There was no significant difference in how much effort participants felt they had to put into understanding each of these two formats in any country.

Participants that saw either of the two frequency formats rated the information as subjectively harder to understand and requiring more effort to understand than those who saw the hazards represented as a percentage (either with or without a graphic), which supports the other part of our hypothesis 4.

In Switzerland, where we had data for subjective perceptions of the combined percentage and frequency format, this combined format was rated intermediate between the percentage (easiest) and frequency (hardest) formats in terms of ease of understanding. This is interesting because subjective comprehension ratings often simply mirror the amount of text involved (with formats with more text typically being rated harder to understand). The combined format has more words on it than any other, and so its relatively high subjective comprehension ratings suggest that the addition of the percentage information to the frequency format actually enhanced comprehension.

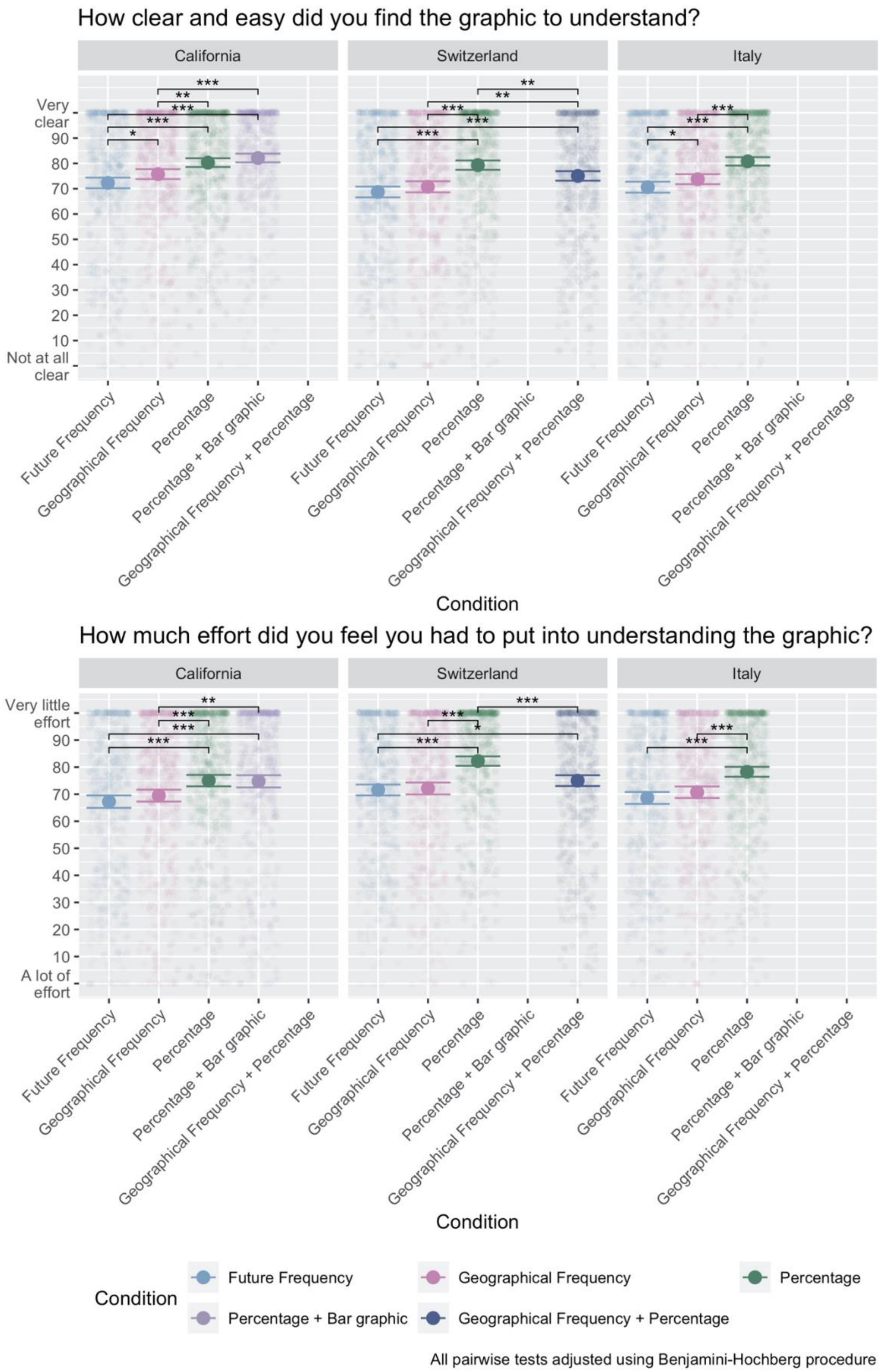


Figure 4: Participants' ratings of how clear they found the information and how much effort they had to put into understanding it, for each format in California (n=2290), Switzerland (n=2383) and Italy (n=2326). Note – not all formats were tested in all countries, hence the blank columns. Individual data points are plotted as jittered points, with means shown as dots with confidence intervals (+/- 95% CIs). Asterisks denote significant pairwise differences between formats, adjusted using the Benjamini-Hochberg procedure (p <.05; ** p <.01, *** p <.001). [N.B. Only three formats tested in Italy due to a mistake with the graphic in one arm at the 22% hazard level]*

Worry

Figure 5 shows participants' ratings of how worried they would be to see a forecast of a 22% chance of a damaging earthquake in each of the formats tested, and for each of the three different countries in turn. This shows that worry mirrors their perception of the risk, as expected – there were no additionally 'worrying' features of any of the formats.

Our third hypothesis (preregistered for Switzerland and Italy) hypothesised that participants seeing the hazards represented as expected frequencies would feel more worried about the forecast than those who saw them represented as percentages (similar to H2 which hypothesised the same relationship but with risk perception as the dependent variable of interest). This hypothesis appears to be confirmed in Italy for both 'future' and 'geographical' frequency formats, which were perceived as significantly more worrying than the percentage format. This significant difference between frequency and percentage formats was only detected for the difference between the 'geographical' frequency format and the percentage format in Switzerland, however.

Although we didn't preregister this hypothesis for California, the data show similar patterns to Italy – with both the future and geographical frequency format rated as more worrying than the percentage format (and also compared to the percentage plus bar graphic format that was used as one of the experimental arms in California).

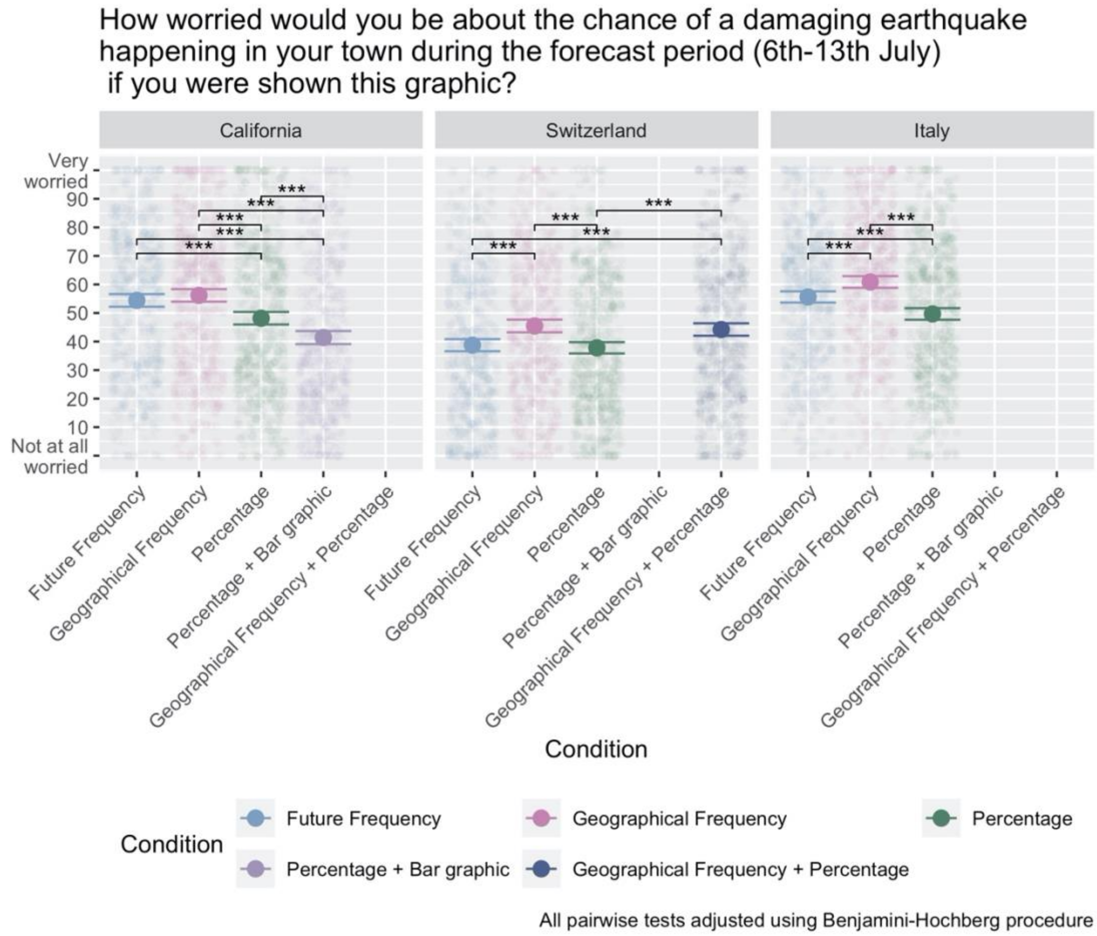


Figure 5: Participants' ratings of how worried they would be to see a forecast of a 22% chance of a damaging earthquake in each format in California (n=2290), Switzerland (n=2383) and Italy (n=2326). Note – not all formats were tested in all countries, hence the blank columns. Individual data points are plotted as jittered points, with means shown as dots with confidence intervals (+/- 95% CIs). Asterisks denote significant pairwise differences between formats, adjusted using the Benjamini-Hochberg procedure (p <.05; ** p <.01, *** p <.001). [N.B. Only three formats tested in Italy due to a mistake with the graphic in one arm at the 22% hazard level]*

3.4 Conclusions

The higher risk perception of the low probability seismic hazards when presented in a frequency format versus a percentage format is stark, and consistent across the three languages tested.

Between the two frequency formats tested, the 'geographical' format that expressed the frequency in terms of *'out of 100,000 towns with exactly the same chance of an earthquake as your town'* often produced a higher perception of the same risk than the more abstract 'future' format using the phrasing *'imagine 100,000 possible ways in which the week could turn out in your town'*. These findings are in line with previous studies of frequencies versus percentages (particularly with high denominators) such as (Peters, Hart and Fraenkel, 2011; Freeman *et al.*, 2021) and psychological theories behind the perception of frequency formats, which suggest that the concreteness of the image conjured by the format increases the perception of the risk (Siegrist, 1997; Slovic, Monahan and MacGregor, 2000; Keller, Siegrist and Gutscher, 2006). Thus, a frequency format that uses a more easily imagined object such as a town (as in the 'geographical' frequency format we tested) should be perceived as riskier than one using a more abstract object such as a 'possible future' (as in the 'future' frequency format we tested), and both should be perceived as riskier than an even more abstract percentage format.

At high hazard levels (22% and 44% in this experiment), the difference in perception of the risk between the frequency and percentage formats decreases at the 44% level (in California there are no significant differences at this level, and in Italy it is only the future frequency format that is significantly lower in perception from the others at the 44% level). This means that the difference in perception between the higher and lower hazard levels is actually greatest in the percentage format. However, as Figure 3 shows, the first three hazard levels (1% and below) are relatively difficult for people to discriminate in the percentage only format (green), whereas the combination of geographical frequency and percentage (dark blue) gives a bigger difference in perception between the first three hazard levels, and so appears to aid discrimination.

Percentage alone is the format rated as clearest and easiest by the audiences – although this doesn't necessarily mean that it gives them all the information that they need to make informed decisions. However, an institution providing "only" a percentage, even with several decimal places, are not necessarily serving their audiences poorly.

To discover whether those seeing the percentage format (with the lowest perceived risk at the lower hazard levels) or those seeing the geographical frequency format (with the highest perceived risk at the lowest hazard levels) – or those seeing both (with an intermediate perceived risk) have a more 'appropriate' perception of the risk requires an analysis of behavioural intentions after seeing this hazard level. This is addressed in Experiment 2.

The difference overall in the perception of the risks between Switzerland (the lowest of the three countries in terms of seismic hazard) and California and Italy (both with a high seismic hazard) is interesting: those in Switzerland perceive the risks as lower across-the-board, and are less worried by them. This is explored further in the analysis of Experiment 3.

4. Experiment 2: Evaluating the best way to give meaningful context to the chance of an earthquake

4.1. Introduction

A single probability is very difficult to evaluate without context or experience. Is 10% big or small? Should you be worried or not?

Members of the public, infrastructure managers, policymakers - all would use operational earthquake forecast probabilities to make decisions such as whether to practice an emergency drill, whether to cancel leave of emergency workers, whether to shut tunnels or bridges. But in order to make such decisions, they need to be able to interpret the probability presented to them (as well as the potential impact of the event should it happen, but here we are talking only about communicating the probability).

Many risk communicators advocate the use of comparator risks to help put an unfamiliar risk in the context of others that might be more familiar (Fischhoff *et al.*, 1978; Wilson and Crouch, 1987; Covello, Sandman and Slovic, 1988; Kunreuther, Novemsky and Kahneman, 2001; Keller, Siegrist and Gutscher, 2006; Pighin *et al.*, 2013). These comparator risks have been described as producing a “conceptual yardstick” (Covello, 1991). Savadori and colleagues (Savadori *et al.*, 2022) found that comparator risks increased the sensitivity of UK participants’ perception of different levels of risk when studying seismic risk communication.

A graphical aid called a risk ladder, which shows the positions of comparator risks along a visual scale, is also often used to communicate unfamiliar risks in a range of fields from medicine to natural hazards. Evidence suggests that people are often good at intuitively estimating relative risks, even if the probabilities they attach to each specific risk can vary a lot (Persoskie and Downs, 2015). Indeed when people are asked to judge a variety of risks on different scales, the same ordering between the different risks often emerges, even if the spacing between them is very different (Fischhoff and MacGregor, 1983). Risk ladders attempt to make use of this consistency in relative risk judgements by helping position a new and unfamiliar risk amongst other risks that people may already have a mental scale for, both relatively and also – sometimes – on an absolute, linear scale as well (Persoskie and Downs, 2015). The visual position of a risk on the scale can help people understand the magnitude and hence importance of the risk (Sandman, Weinstein and Miller, 1994; Siegrist, Orlow and Keller, 2008).

The selection of comparator risk, and thus where the risk to be communicated sits on the visual risk ladder scale, are likely to influence people’s perception of the likelihood of that risk. By choosing risks that are substantially more likely than the low likelihood risk, it is possible to minimise people’s perception of the risk, just as choosing many low likelihood risks, the majority of which are lower in likelihood than the risk to be communicated, can enhance people’s perception of the risk (Sandman, Weinstein and Miller, 1994; Siegrist, Orlow and Keller, 2008).

The simplest context, however, is a single comparator, such as a baseline. How much a risk is elevated above the baseline – expressed either in absolute or relative terms – might be enough of a context cue, and less confusing to an audience, than a series of comparators. However, giving the relationship of a risk with its baseline in a relative risk format alone, although probably easy to comprehend, can very often make a difference from this baseline seem very large, particularly when the absolute baseline hazard is very small. Giving people the baseline hazard as an absolute risk in addition to the current hazard level as an absolute risk is likely to make hazard levels with a small absolute difference from baseline look less hazardous than when expressed as a relative risk – but a single comparator may not be enough to give people a sense of context (Kunreuther, Novemsky and Kahneman, 2001).

Previous authors e.g. (Kunreuther, Novemsky and Kahneman, 2001) have suggested that comparators that give more of a scale than a simple baseline, that are more vivid in the images or narratives they create, that are easily comparable to each other, and that are more familiar to the readers, are more useful in providing sensitivity to differences in low probabilities. Similar to Kunreuther et al. (2001), Freeman et al. (Freeman *et al.*, 2021) also found, through interviews with members of the public, that comparators that were 'the same risk' (in this case the risk of death from COVID-19) but for other people that faced different hazard levels (for reasons that were easily understood, such as age) were considered more helpful than comparators that were 'other risks' (e.g. the risk of death from other causes). Translating these experiences to the seismic domain would suggest that the most useful comparators for seismic hazard in one area would be comparison with the seismic hazard in other areas that the audience would have a sense of.

In this experiment, therefore, we are attempting to identify the effects of adding different kinds of context to a forecast absolute seismic hazard level: a relative risk (compared to baseline); the absolute baseline hazard level; a risk ladder showing the normal seismic hazard level of familiar comparator cities; or the same risk ladder but with the comparator cities replaced with 'other risks' as comparators, in the form of other causes of death (similar to Savadori *et al.*, 2022).

The actual numbers (and hence positions along the ladder) assigned to the risks on the two risk ladders (city comparators or other comparators) were identical, so that all that we were comparing in the experiment was the effect of the types of risks labelled on them. Cities and 'other risks' were chosen to have approximately their true hazard level illustrated so as not to introduce perceptual incongruity for participants. Before the experiment, however, participants were asked to rate the risks we used as comparators so that we could assess whether their perception of the relative risks matched each risk's position on the risk ladder or not.

The 'other risks' were chosen to be serious risks (death), and were kept constant across different countries (but said to refer to the likelihood of death by that means in the country of the participant). The cities were chosen such that there were two cities within the country of testing (representing a low and a high hazard for that country), and two international ones.

In this experiment the control condition is the absolute risk alone, expressed as a percentage. We do not test the effects of combining another contextual element with other numerical formats, such as the two frequency formats in Experiment 1.

4.2. Methods

Before the experimental portion of the survey, we asked participants to rate how high they thought the earthquake risk was in a range of international cities, including those used in the risk ladder portraying the seismic risk in different cities. Participants in California were given the option to opt out of rating a city they had not heard of, but participants in the other countries were asked to give their best guess. We also asked them to rate the 'other risks' used in the risk ladder portraying non-seismic risks.

Participants were then randomised to one of four arms, and each participant saw either three (in California) or four (in Italy and Switzerland) stimuli in a randomised order within their respective arm, each showing a different hazard level: 0.003%, 2.1%, (3.9%), 6.7%.

The four arms were chosen from a set of five possibilities, with different comparisons made in different countries (in order to keep an appropriate level of power we could have no more than four arms at any one time). See Figure 6 for the graphics used in each arm in each country.

In Italy, in order to include all five stimuli for direct comparison with each other, after Experiment 2, participants were further (independently) randomised to view a single extra stimulus in the relative risk condition in one of the four hazard levels, and asked the risk perception and preparedness questions that served as two of our dependent variables (see below).

Arm 1 – Absolute risk
Absolute current risk only (shown as a percentage)

Arm 2a – Relative risk ('x times higher')
Absolute and relative current risk (how many times higher the current risk is than baseline) (California only, although included in Italy as an extra within-subjects question)

Arm 2b – Baseline risk
Absolute current risk and baseline current risk (Switzerland and Italy only)

Arm 3 – 'City comparators'
Absolute current risk shown on a risk ladder with comparators of seismic hazard in other cities (two from the country of testing, and two international)

Arm 4 – 'Other comparators'
Absolute current risk shown on a risk ladder with comparators of other risks

We collected data on six main dependent variables:

1) **Risk perception**

Imagine that you visited a website that showed you this graphic indicating the chance of a damaging earthquake happening in your town within the next 7 days. How would you classify that risk in your mind?

Move the slider below to indicate it: 'Very low risk' – 'Very high risk'

2) **Subjective comprehension** (measured once only per arm, at the 2.1% hazard level), measured via two questions:

How much effort did you feel you had to put into understanding the graphic?

Move the slider below to indicate it: 'A lot of effort' - 'Very little effort'

How clear and easy did you find the graphic to understand?

Move the slider below to indicate it: 'Not at all clear' – 'Very clear'

3) **Worry**

How worried would you be about the chance of a damaging earthquake happening in your town during the forecast period (6th-13th July) if you were shown this graphic?

Move the slider below to indicate it: 'Not at all worried' – 'Very worried'

4) **Surprise** (understanding of uncertainty):

How surprised would you be if a damaging earthquake happened in your town during the forecast period (6th-13th July) if you had seen this graphic?

Move the slider below to indicate it: 'Not at all surprised' – 'Very surprised'

5) **Preparedness action intentions**:

Looking again at this earthquake forecast, imagine it was for the next 7 days in your town. Which of the below actions would you take in response? Please choose all that apply.

- Actively look online for information on how to prepare for or respond to earthquakes.
- Practise an earthquake drill
- Prepare and pack a bag with essential items specifically for if an earthquake happens (e.g. water, clothes, ID, medicines).
- Increase the earthquake resistance of your residence (e.g. strengthen the building, secure its foundations, or secure furniture such as bookcases)
- Prepare a household emergency plan for an earthquake to ensure you and your family/children know what to do
- Practise your household emergency plan for an earthquake to ensure you and your family/children know what to do
- Undertake First Aid training specifically to assist after an earthquake
- Immediately call nearby family and friends to tell them about the forecast
- Put supplies (e.g. tent, blankets, clothes, water, canned food) in your car in case an earthquake makes your house uninhabitable
- Prop doors with door stops (excluding fire or security doors) to make it quicker to get out of the building
- Go to stay with friends or family in another part of the country for the week
- Sleep outside in an area where no buildings could fall on you
- Do nothing in response

6) **Usefulness of information** (measured once only per arm, at the 2.1% hazard level. 'Useful' question only asked in California):

How useful did you find this graphic for informing you about the chance of an earthquake happening in your town?

Move the slider below to indicate it: 'Not at all useful' – 'Very useful'

How satisfied are you that the graphic adequately informed you about the chance of an earthquake happening in your town?

Move the slider below to indicate it: 'Not at all satisfied' – 'Very satisfied'

In addition, in Switzerland and Italy, at the end of the experiment we showed participants all five formats and asked them to rank them in usefulness and give us free text reasons why they had ranked them the way they did.

We additionally collected potential covariates including education, objective numeracy, optimism/pessimism, fatalism and earthquake experience.

We pre-registered a series of formal hypotheses at https://osf.io/5kf9r/:

1. Participants in all arms seeing three/four stimuli representing different hazard levels will perceive the risks to be higher in the stimuli showing the higher hazard levels.

2. a) For hazard levels above baseline, participants seeing hazard represented as a relative risk ('x times higher': arm 2 in California, additional within subjects arm in Italy) will perceive the hazards to be riskier than those who see the absolute risk alone with no comparators (arm 1). (Preregistered in California only.)

   b) For hazard levels above baseline, participants seeing the baseline hazard represented alongside the current absolute risk level (arm 2 in Switzerland and Italy) will perceive the hazards to be riskier than those who see the current absolute risk level alone with no comparators (arm 1) (Preregistered in Switzerland and Italy only).

3. Participants being shown a risk ladder (arms 3 and 4) will be more sensitive to the differences between the three/four hazard levels (i.e. perceive the higher levels as presenting higher risks that are more distinct from one another and from the baseline hazard level) than participants shown the absolute risks only (arm 1).

4. Participants shown a risk ladder (arms 3 and 4) will rate the information as harder to understand than those not shown a risk ladder (arms 1 and 2). Note that this will only be tested for the one hazard level where the subjective comprehension questions are asked i.e. 2.1%.

5. Participants shown a risk ladder (arms 3 and 4) will rate the information as more useful than those who were shown a relative risk (arm 2 in California, additional within subjects arm in Italy) and they will all rate it as more useful than those shown only an absolute risk (arm 1). Note that this will only be tested for the one hazard level where the usefulness questions are asked i.e. 2.1% (Preregistered in California only).

6. Participants shown both the absolute current risk level and the baseline hazard (arm 2 in Switzerland and Italy) will rate themselves more satisfied by the information in the format than those shown only the absolute current risk level (arm 1). Note that this will only be tested for the one hazard level where the satisfaction question is asked i.e. 2.1%. (Preregistered in Switzerland only).

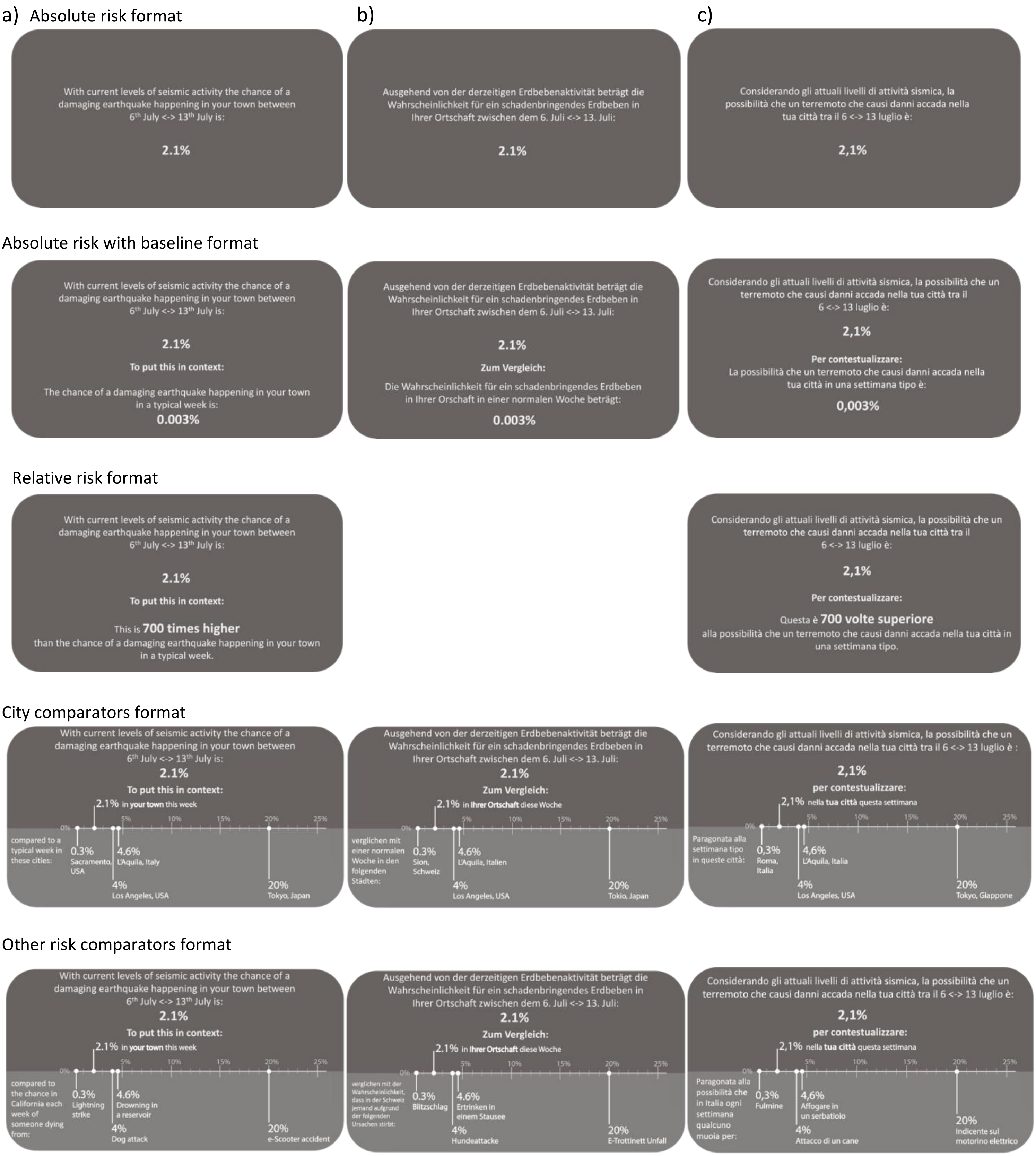


Figure 6: Graphics used in Experiment 2 in a) California (the combined absolute risk with baseline format was not tested here but is included for the English translation it provides), b) Switzerland and c) Italy (the relative risk format was included as a separate, within subjects question rather than a fifth between subjects arm tested directly against the rest).

4.3. Results

Risk perception of different comparators

Figure 7 shows how participants in different countries rated the baseline seismic hazard in a range of different cities (participants in each country were given at least three cities in their own country, including those used in the risk ladder, plus a range of international cities – also including the two used in the risk ladder). Participants in California were given the option to opt out of rating a city they had not heard of, but participants in the other countries were asked to give their best guess, and many left the slider at its anchor point of 50%, which is clear in the results.

It is striking that participants across all three countries had similar perceptions of the seismic hazard in each city, not just in relative terms (to each other) but in absolute terms (e.g. Swiss participants did not rate the seismic hazard much lower across the board as we see in their perceptions of the probabilities in each experiment). The biggest exception is L'Aquila, in Italy, which participants in Italy perceived as a high hazard, likely due to the recent and destructive earthquake there in 2009. Participants outside Italy were probably much less familiar with the town and the 2009 earthquake, which explains their lower rating of the hazard level there (although participants in Switzerland may be more familiar than those in California, looking at the comparative ratings).

L'Aquila was one of the exemplars that was used as a comparator on the seismic comparator risk ladder across all three countries, and the different perception of the seismic hazard in the town may have influenced different reactions to the risk ladder between countries.

Apart from that town, participants correctly identified the relative risks (rankings) of the other comparator cities used in the risk ladder they were presented with, although these are very difficult to calculate (given the varying areas of the different cities etc).

Figure 8 shows the ratings of participants in different countries of their perception of the likelihood of four non-seismic risks, later used as comparators on the second risk ladder we tested. Participants in California correctly identified the order of the risks relative to each other, except in placing the risk of drowning in a reservoir above that of the e-scooter risk. Participants in Italy and Switzerland perceived the risks differently from Californians, but the same as each other in terms of how they ranked them relatively. They correctly identified the e-scooter risk as the highest.

It is notable how low participants rated these risks compared to the relative hazards of the cities. Although asked on a different page of the survey, so not directly comparable, if participants genuinely considered these risks lower than the seismic risks it may affect their perceptions of risks positioned against these on a risk ladder.

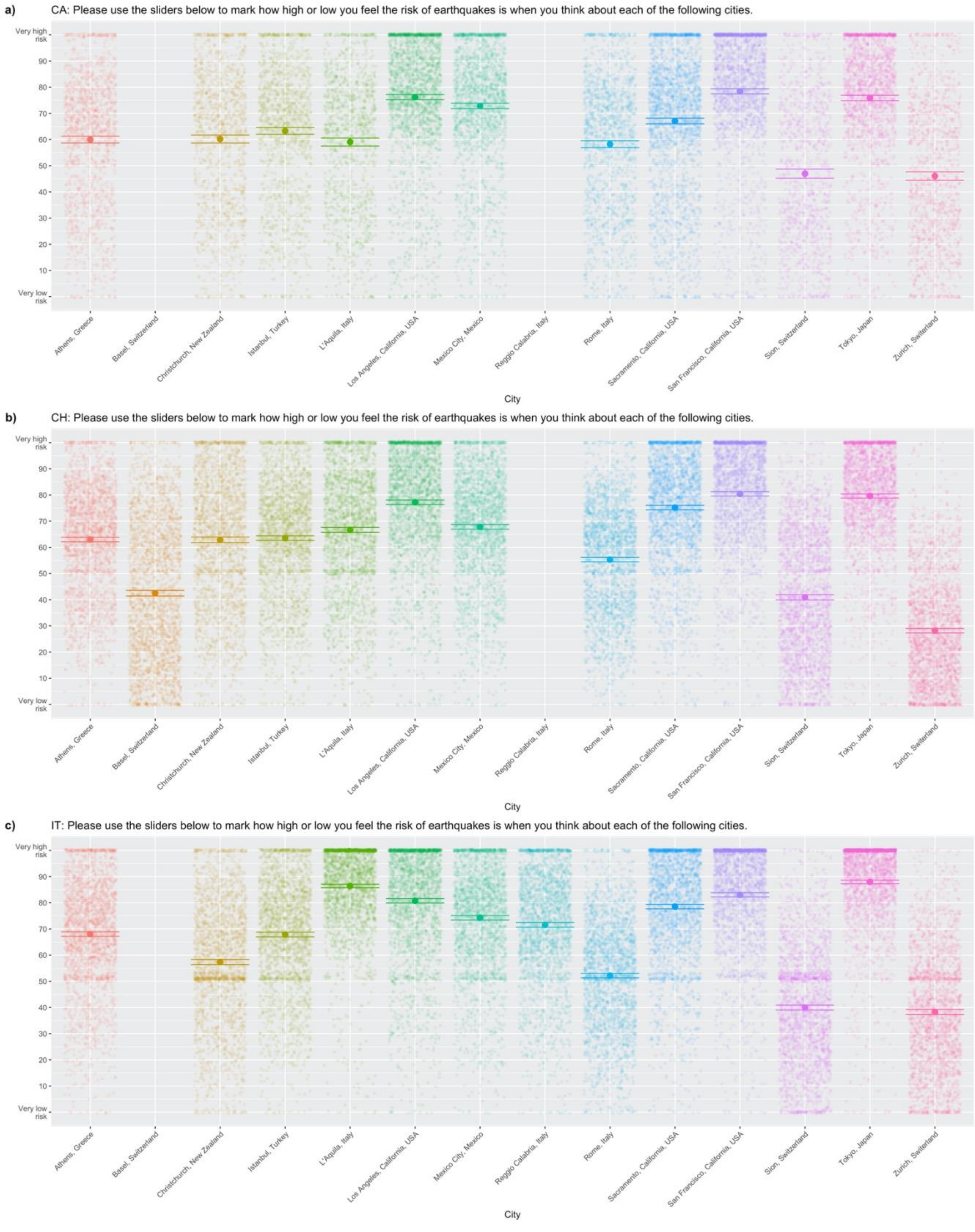


Figure 7: Participants' ratings of their perceived risk of an earthquake in different cities, participants from a) California (n=2290), b) Switzerland (n=2383) and c) Italy (n=2326). Individual data points are plotted as jittered points, with means shown as dots with confidence intervals (+/- 95% CIs).

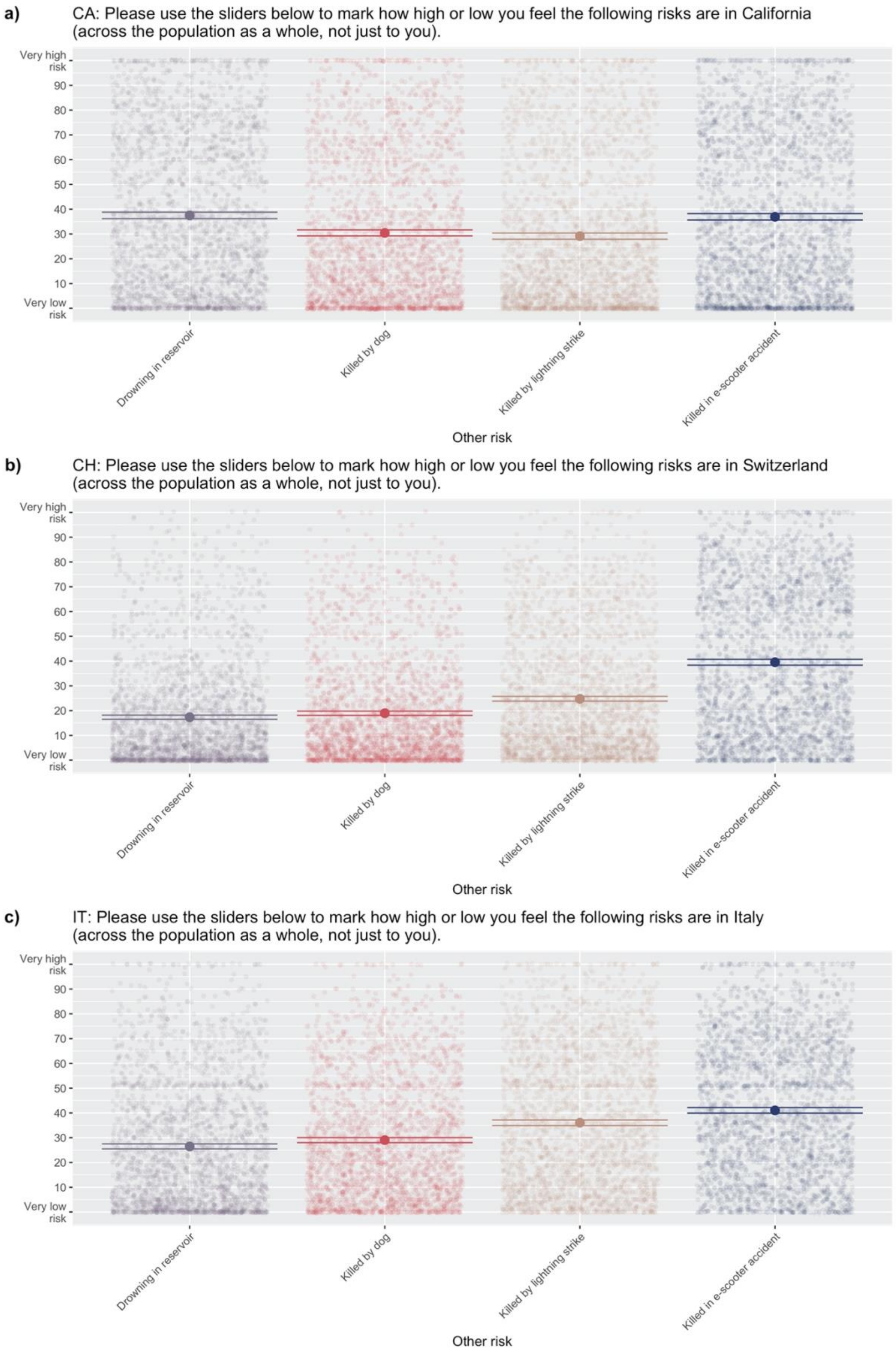


Figure 8: Participants' ratings of the risks of different events happening to people in their country, participants in a) California (n=2290), b) Switzerland (n=2383), c) Italy (n=2326). Individual data points are plotted as jittered points, with means shown as dots with confidence intervals (+/- 95% CIs).

Risk Perception of different hazards

Figure 9 shows participants' ratings of how risky each of the hazard levels portrayed (within-subjects) felt in each of the different formats (between-subjects). From these results it's clear that our first two hypotheses are supported. Firstly, participants in all arms seeing stimuli representing different hazard levels perceived the risks to be higher in the stimuli showing the higher hazard levels.

Secondly, for hazard levels above baseline, participants seeing hazard represented as both an absolute and a relative risk together ('x times higher': arm 2 in California and given as an extra between subjects question in Italy) perceived the hazards to be riskier than those who saw the absolute risk alone with no comparators (arm 1). In fact, in California this relationship even held for the representation of 0.003%, which was the baseline hazard itself (remember that in California, the presentation of the different hazard levels was a within-subjects experiment where these different hazard stimuli were shown in a random order: previous studies have shown that there is an overall perceptual effect of seeing multiple hazards presented in a format which heightens risk perception, such as a relative risk or expected frequency (Freeman *et al.*, 2021)).

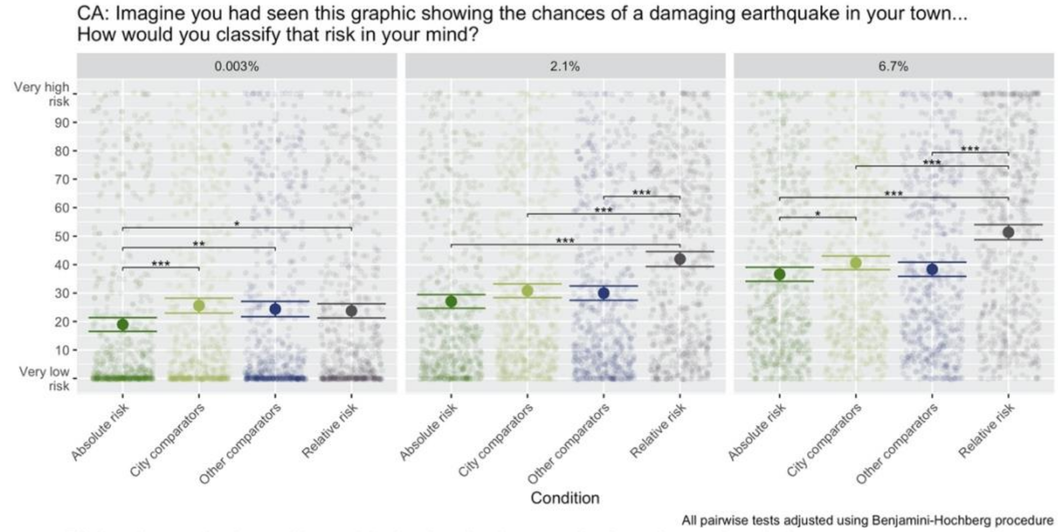
However, our third hypothesis, that participants seeing one of the two risk ladder formats would be more sensitive to differences between the different hazard levels than those seeing the absolute risk only, was not supported. Instead, it looks as though the seismic city comparator risk ladder, at least, increases the perception of the risk (compared with the absolute risk alone) by a fairly constant amount across all hazard levels. Further statistical analysis will be needed to clarify this effect.

It is notable that seeing an absolute risk percentage alongside the seismic city comparator risk ladder resulted in participants having a higher perception of that risk than if it were presented alongside a risk ladder with other (non-seismic) risks on it. This may well relate to participants' perceptions of the non-seismic risks, which we elicited earlier in the survey. People generally perceived the risks of various fatal accidents as low compared with the risks of an earthquake in the cities used on the seismic city comparator risk ladder. This may explain the difference in perception created by the two risk ladders.

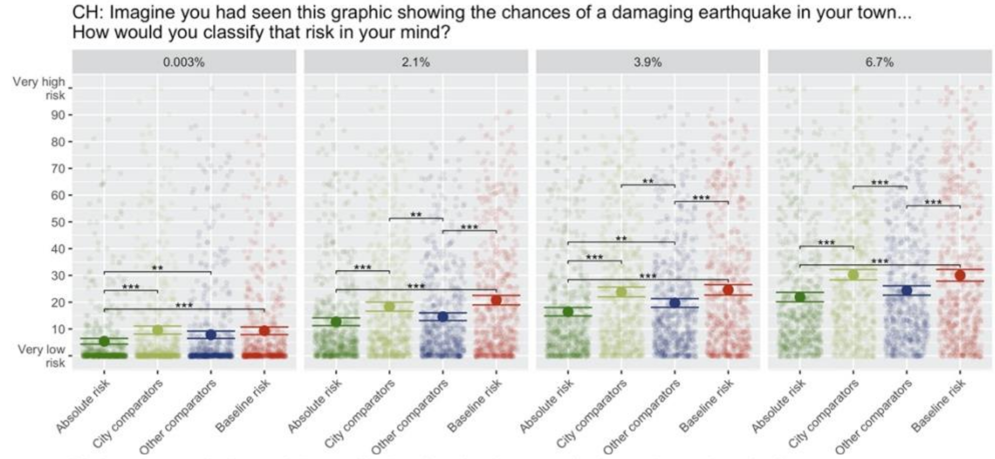
Our results are also only partly in support of our hypothesis 2b. In Switzerland, the combined format of absolute current risk alongside the baseline hazard, was perceived as riskier than the absolute current risk alone at hazard levels above baseline (stated to be 0.003%), but also at the hazard level where the current risk equalled the baseline hazard. Furthermore, in the highest hazard level (6.7%) in Italy, the relationship was inverted; those seeing the absolute current and baseline risks together perceived the risk as lower than those seeing the absolute current risk alone.

Perhaps the most interesting result regarding this hypothesis, however, was that (as in Experiment 1) we see that the perception of the absolute risks alone in Switzerland is lower (across all hazard levels) than that in California or Italy. However, when shown the absolute risk in comparison to the baseline hazard, the perceptions of each hazard level in this format in Switzerland almost matched that in Italy. In other words, when given the same baseline hazard, people in Switzerland and Italy responded similarly, but without being given a baseline hazard, those in Switzerland perceived the risks as much lower than those in Italy. It is interesting that the 'other risks' risk ladders (which also gave each country identical comparators) did not have this same, normalising effect.

a)

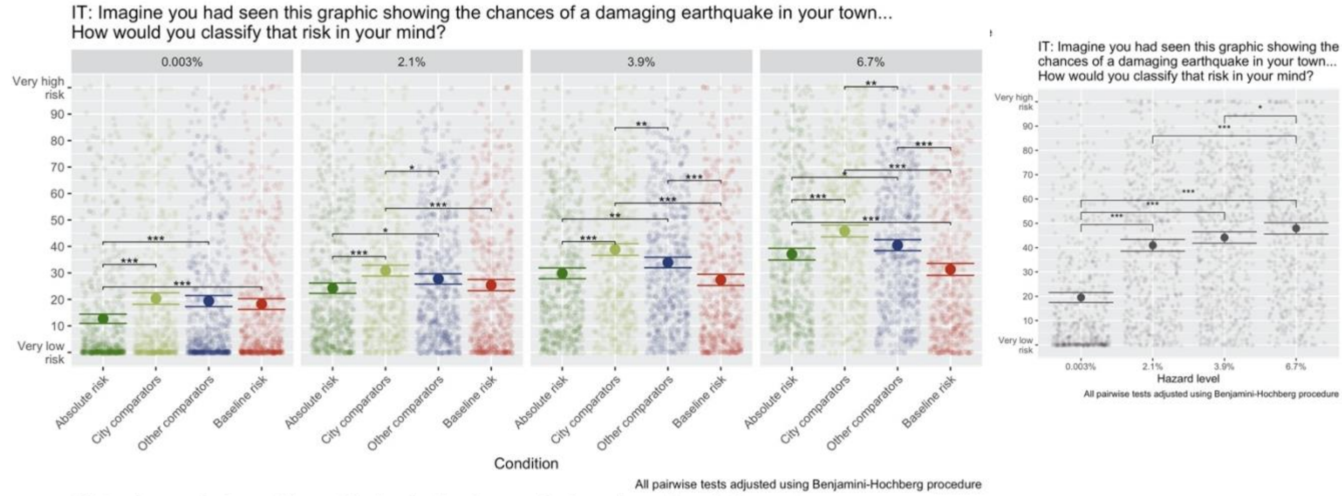


b)



c)

i)



ii)

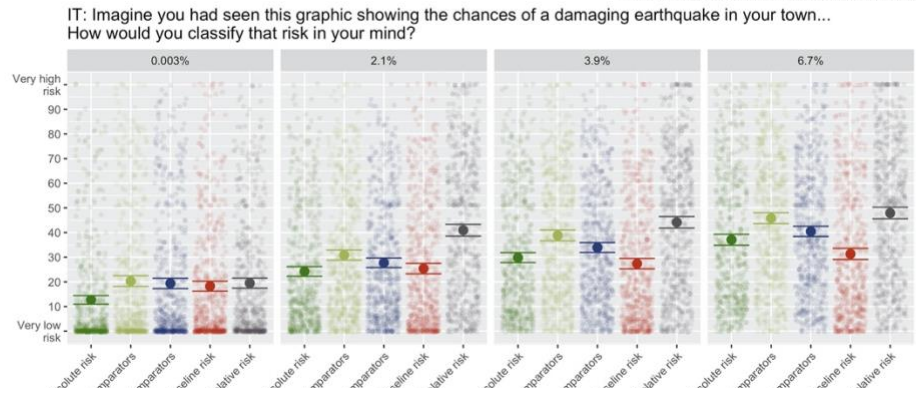


Figure 9: Participants' ratings of their perception of the risk represented by different hazard levels in different formats in a) California (n=2290), b) Switzerland (n=2383) and c) i) Italy (n=2326). In Italy, the Relative Risk arm was presented between subjects, not within subjects like all the others, so shouldn't be compared statistically. However, all five arms are also illustrated within the same graph (ii) for ease of visual comparison. Individual data points are plotted as jittered points, with means shown as dots with confidence intervals (+/- 95% CIs). Asterisks denote significant pairwise differences between formats, adjusted using the Benjamini-Hochberg procedure (p <.05; ** p <.01, *** p <.001).*

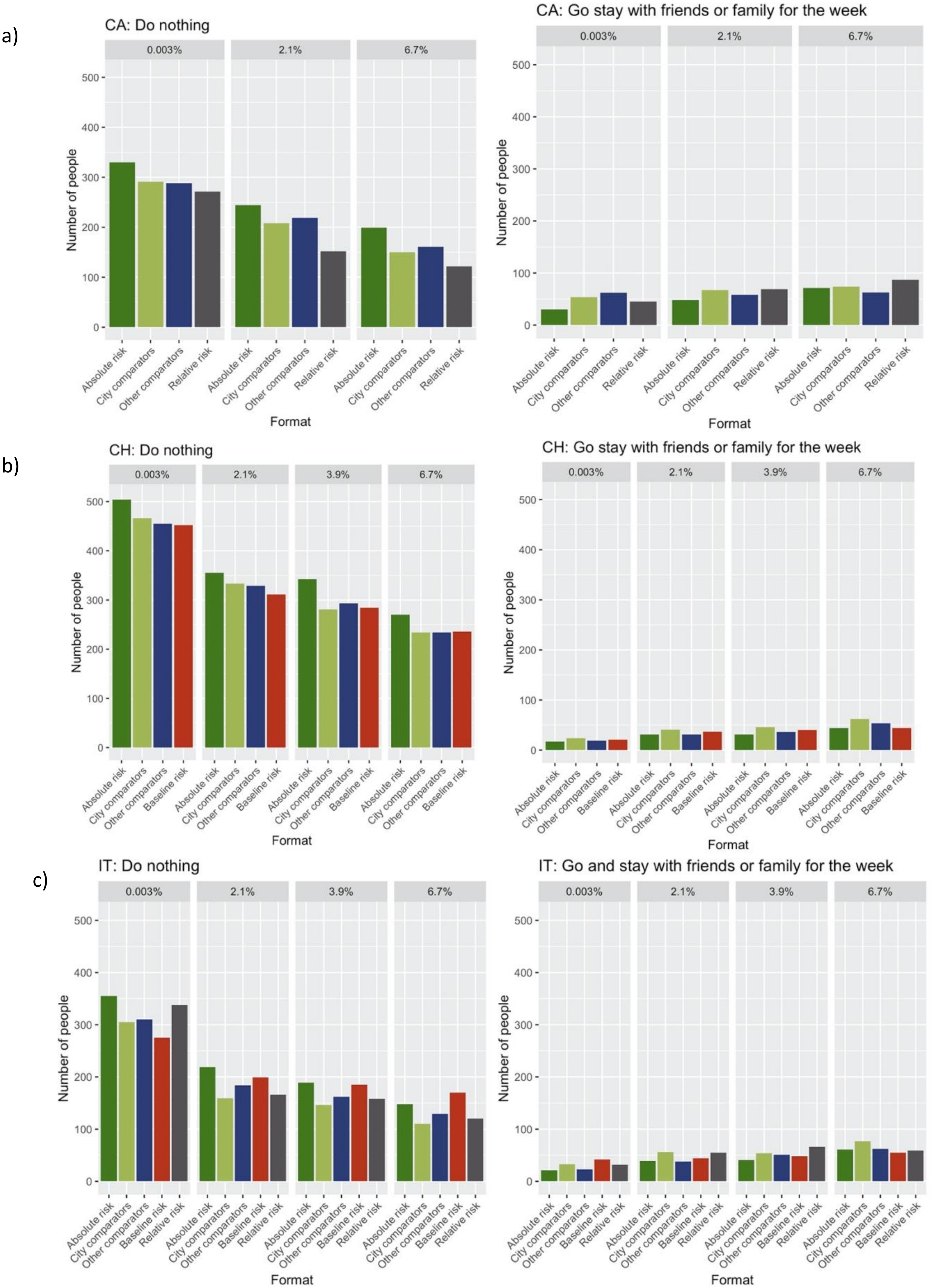


Figure 10: Counts of the numbers of participants who selected that they would 'do nothing' or 'go and stay with friends or family in a different part of the country for the week' (the most extreme responses) in response to seeing an earthquake forecast in their area of a particular hazard level in a particular format in a) California (where the fourth arm was the relative risk format) (n=2290), b) Switzerland (n=2383) and c) Italy (where the fourth arm was the baseline risk format) (n=2326).

Preparation activities

Figure 10 shows how many participants chose the option ‘I would do nothing’ (left hand side) and how many chose the option ‘I would go and stay with friends or family in another part of the country for the week (right hand side), when asked what preparatory actions they would take upon seeing the forecast for their area at different hazard levels, in different formats.

It is impossible to say what is the ‘correct’ perception of a risk to have, but translating perception into actions that people say they would take as a result gives us more of a sense of whether the format is conveying what experts might consider an ‘appropriate’ level of concern.

The shape of each of the ‘do nothing’ response curves largely fit the ‘riskiness’ ratings for each country, with those formats and hazard levels perceived as more risky having fewer participants choosing to ‘do nothing’ in response. They also fit the riskiness ratings between countries, with those in Switzerland apparently being more likely to choose to ‘do nothing’ (i.e. less willing to take action, in line with their lower perceptions of the risk).

The shape of the curves showing the frequency of people choosing the most severe measures (such as leaving the area) broadly mirror the pattern of the risk ratings (with higher perceived risk relating to a higher number of people choosing this extreme option, but they also show that relatively few would take this measure with any of the hazard levels presented.

Still, it is subjective whether these results indicate ‘too much’ or ‘too little’ action in the face of these hazard levels.

Subjective comprehension

Figure 11 shows participants’ subjective feelings about how easy each format was to understand, and relatedly how much effort they felt they had to put into understanding these formats.

These results partially confirm our fourth hypothesis, in that participants rate the two formats showing a risk ladder as harder to understand than the format showing the absolute percentage risk alone without a risk ladder, however participants rated the relative risk and baseline risk as harder to understand than we had anticipated. The risk ladder with the city comparators was thought harder to understand than that with the non-seismic risks as comparators (such as being killed by lightning) – significantly so in Italy and Switzerland.

Subjective usefulness

Figure 12 shows participants’ ratings of how satisfied they were that the graphic adequately informed them of the chance of an earthquake happening in their town, with an additional graph showing Californian participants’ ratings of how useful they found the graphic for informing them about earthquakes (this latter question was only asked in California).

Contrary to our fifth hypothesis, participants in all three countries were less satisfied when they were shown the risk ladder with the city comparators than they were when shown the plain absolute risk level alone (and in California, the relative risk format was also rated as less satisfying than the plain absolute risk level). In Italy, the risk ladder with city comparators was additionally rated as less satisfying than the risk ladder with other risk comparators, and than the format that showed the absolute risk alongside the baseline risk. There were no differences in ratings of usefulness between formats in California. Contrary to our sixth hypothesis, there was no difference in satisfaction levels between participants who saw the baseline risk alongside the absolute risk and those who saw the absolute risk alone.

To investigate these findings further, we added a question in the Swiss and Italian surveys in which people were shown all five formats (due to a survey error, some Swiss participants did not see the relative risk format) and asked to rank them and give us details about their preferences.

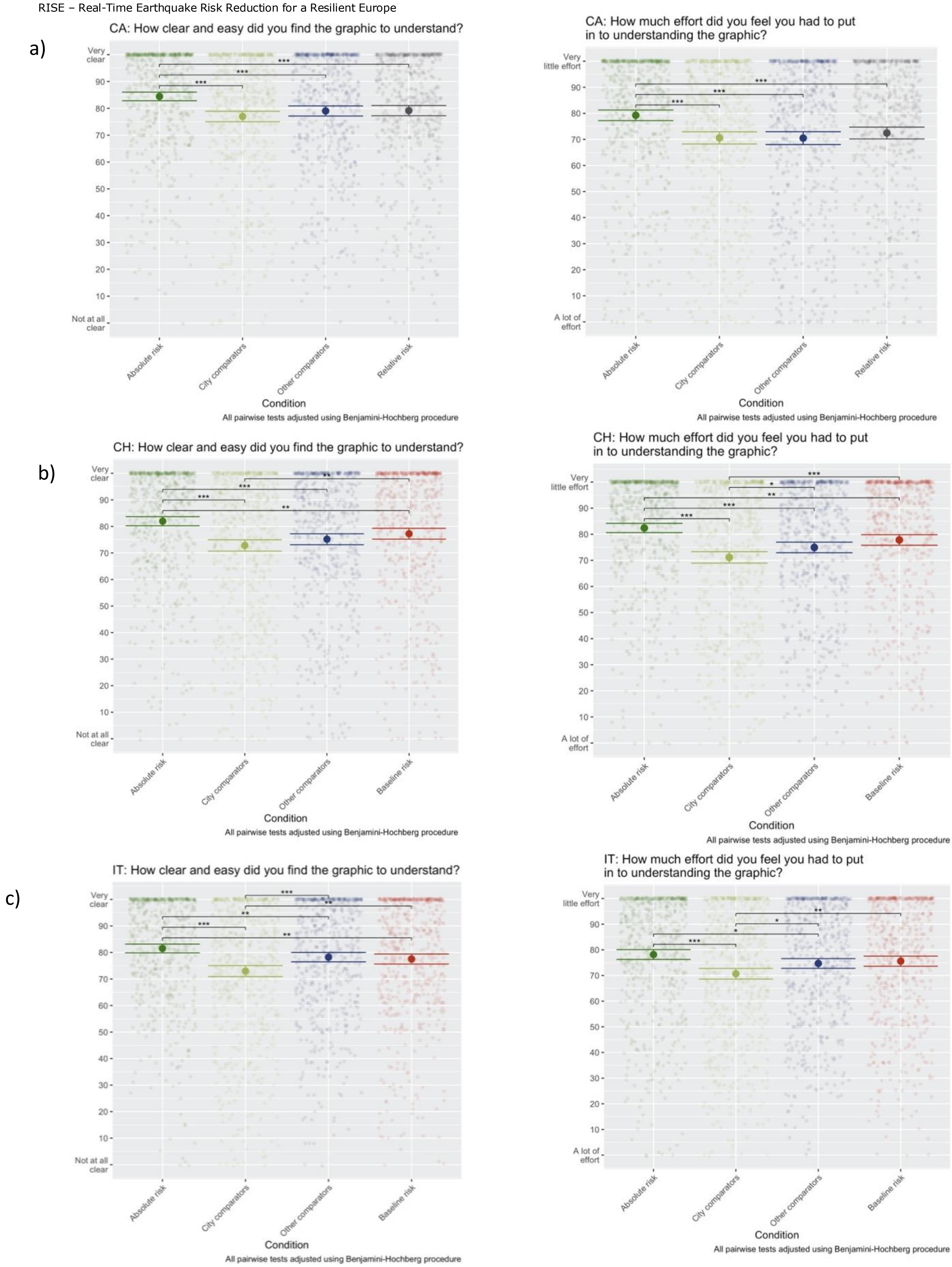
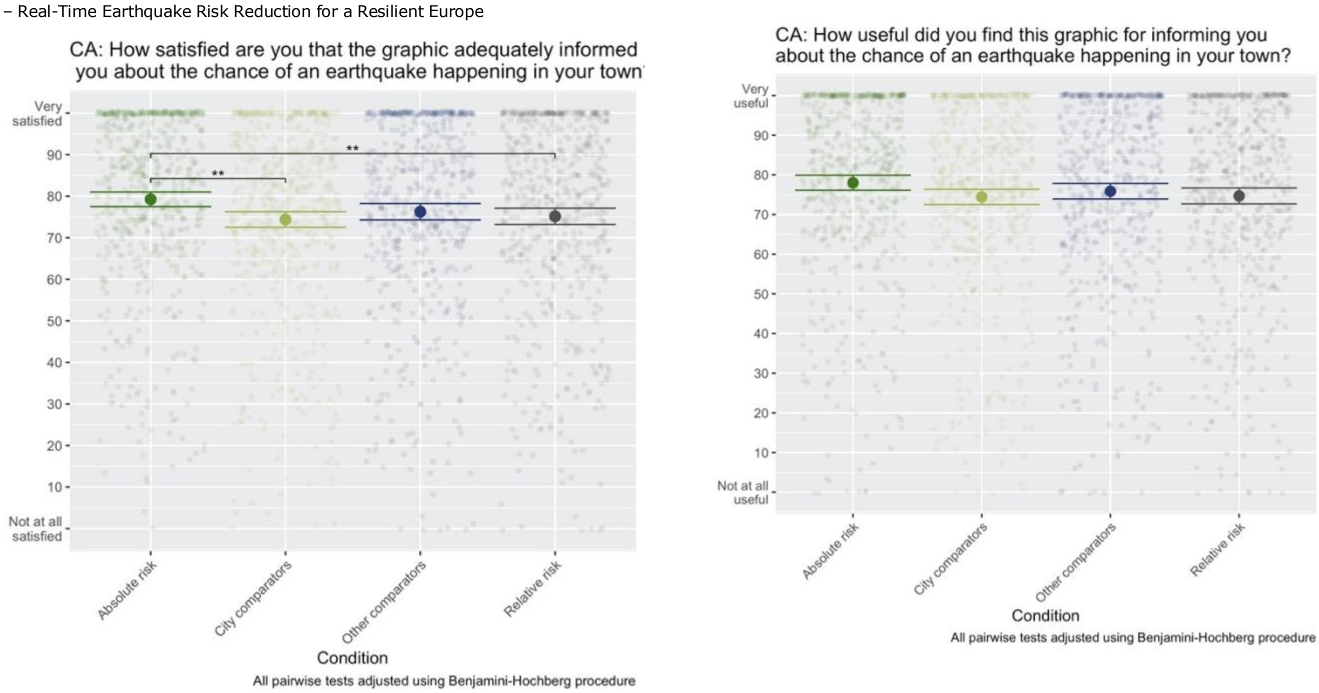
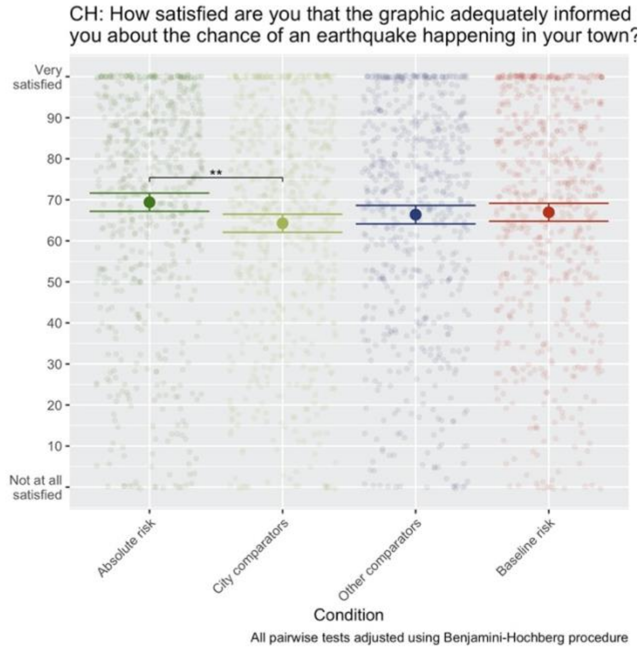


Figure 11: Participants' ratings of how clear they found each format to understand, and how much effort they had to put in to understanding it, in a) California (n=2290), b) Switzerland (n=2383) and c) Italy (n=2326). Individual data points are plotted as jittered points, with means shown as dots with confidence intervals (+/- 95% CIs). Asterisks denote significant pairwise differences between formats, adjusted using the Benjamini-Hochberg procedure (p <.05; ** p <.01, *** p <.001).*

a)



b)



c)

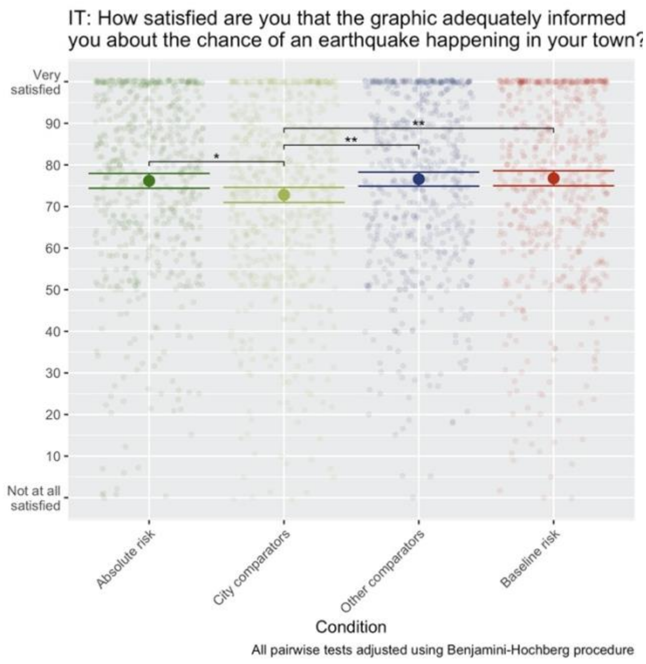


Figure 12: Participants' ratings of how satisfied (and useful in California) they were that the graphic they saw adequately informed them about the chance of an earthquake occurring. a) California (n=2290), b) Switzerland (n=2383), c) Italy (n=2326). Individual data points are plotted as jittered points, with means shown as dots with confidence intervals (+/- 95% CIs). Asterisks denote significant pairwise differences between formats, adjusted using the Benjamini-Hochberg procedure (p <.05; ** p <.01, *** p <.001).*

Subjective preference (ranking)

The free text box gave us insights into people's thinking (although some obviously had difficulty actually ranking the graphics as their individual comments do not match the numerical rankings they gave the graphics). See Figure 13 for rankings.

The text shows a great divergence in preferences and viewpoints. At the heart is the trade-off between simplicity and increased context. Some emphatically ordered them with the 'simplest' and 'easiest to understand' first, and others by the amount of information they contained.

Several mentioned that they wanted a comparator, and many also mentioned a graphic being useful, but then said that they found the risk ladders too complex and didn't want to have to do the maths involved in being shown the absolute baseline and the absolute current risk.

A few also wanted the graphic to emphasise the danger of an earthquake, and some wanted it to reassure.

Some also said that they didn't believe earthquakes were forecastable and this affected their choice: "*I would like as little information as possible, because I believe that an earthquake is simply not really foreseeable*" ("*würde mir möglichst wenige Angaben wünschen, da ich glaube, dass ein Erdbeben einfach nicht wirklich vorauszusehen ist*") (Swiss participant).

For example:

"I think the last two charts are too complicated, as they take too long to be understood correctly. A more concise information I think can reach more people" (Italian participant)

Absolute risk	Baseline hazard comparator	Relative risk	Other risk comparators	City comparators

"The first graphic is the easiest to understand, and does not appear misleading. the last graphic is the one that requires some effort to interpret the figure of 2.1%, and in the end the additional information, regarding the 700 times, is not very significant and can create confusion, uncertainties or doubts about the severity" (Italian participant)

Absolute risk	Other risk comparators	Baseline hazard comparator	City comparators	Relative risk

"Because I ordered them according to their clarity, I would prefer to see a simpler and more immediate graph, less technical" (Italian participant)

Absolute risk	City comparators	Relative risk	Baseline hazard comparator	Other risk comparators

"Only one number is probably the most understandable. A relation to more "everyday" things is also OK. The reference to other places, which are better known with regard to earthquakes, could quickly cause panic. And two percentages confuse." (Swiss participant)

Absolute risk	Other risk comparators	City comparators	Baseline hazard comparator

"The order is based on the graphics that in my opinion better contextualize the risk avoiding excessive alarmism." (Italian participant)

Baseline hazard comparator	City comparators	Absolute risk	Other risk comparators	Relative risk

"Because the 1st is the easiest to understand, has a comparison. The 2nd just a number. The others are a bit more difficult to understand." (Swiss participant)

Baseline hazard comparator	Absolute risk	City comparators	Other risk comparators

"Comparison is good. You can get a better picture. But only earthquake comparison. One should not compare apples with pears" (Swiss participant)

City comparators	Absolute risk	Baseline hazard comparator	Other risk comparators

"I find it especially interesting if you have a comparison. Because otherwise you don't really know how to classify the numbers" (Swiss participant)

Other risk comparators	Absolute risk	City comparators	Baseline hazard comparator

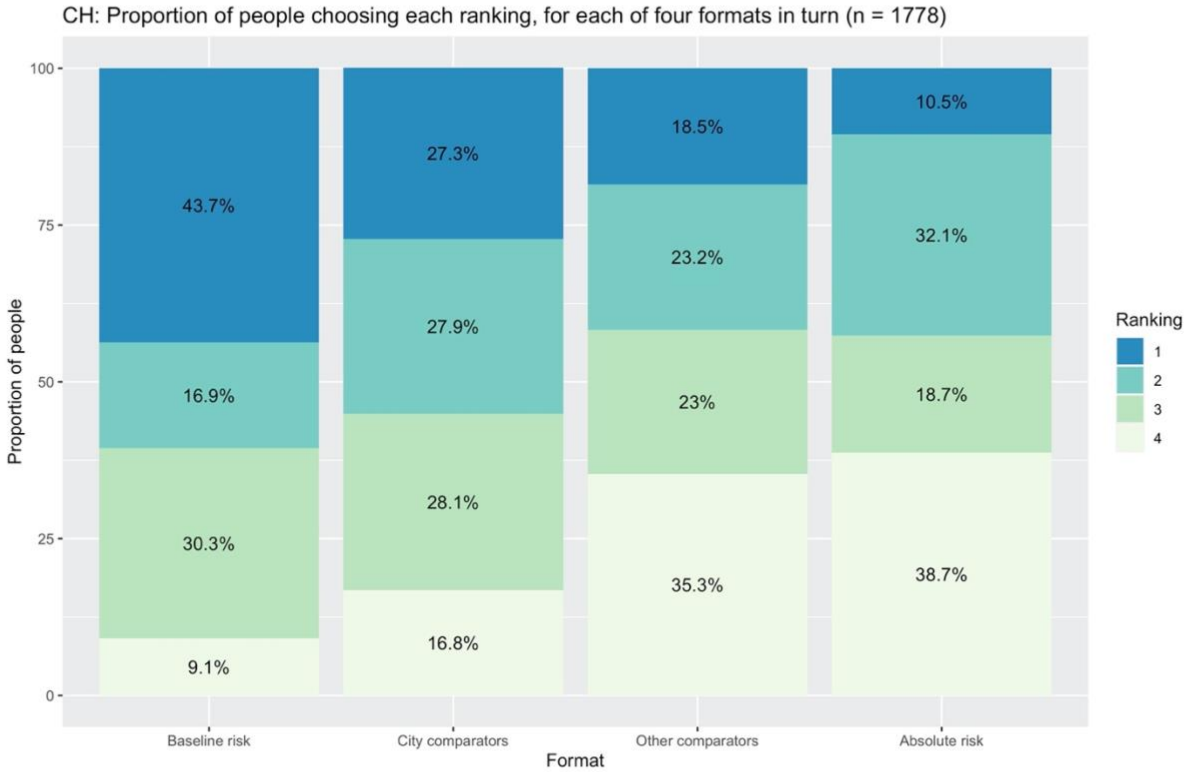
"Even if it takes a moment to understand the graphics, the top one is the most useful, because you get very detailed information and what you can imagine under it." (Swiss participant)

Other risk comparators	City comparators	Absolute risk	Baseline hazard comparator

The rankings echo this difference in preference between comparator information and simplicity. In Switzerland, some form of comparator was preferred by the majority, but here, the format with the local absolute baseline risk level as a comparator was ranked top, above the risk ladder with seismic city comparators.

In Italy, the risk ladder with the seismic city comparators was preferred and the local absolute baseline risk level comparator was the bottom ranked.

The format with the local absolute baseline risk performs very differently in Switzerland and Italy, and this may be due to language differences or cultural differences in these two countries. The seismic city comparator risk ladder was ranked either first or second in the two countries so is consistently liked by a large proportion of the population.



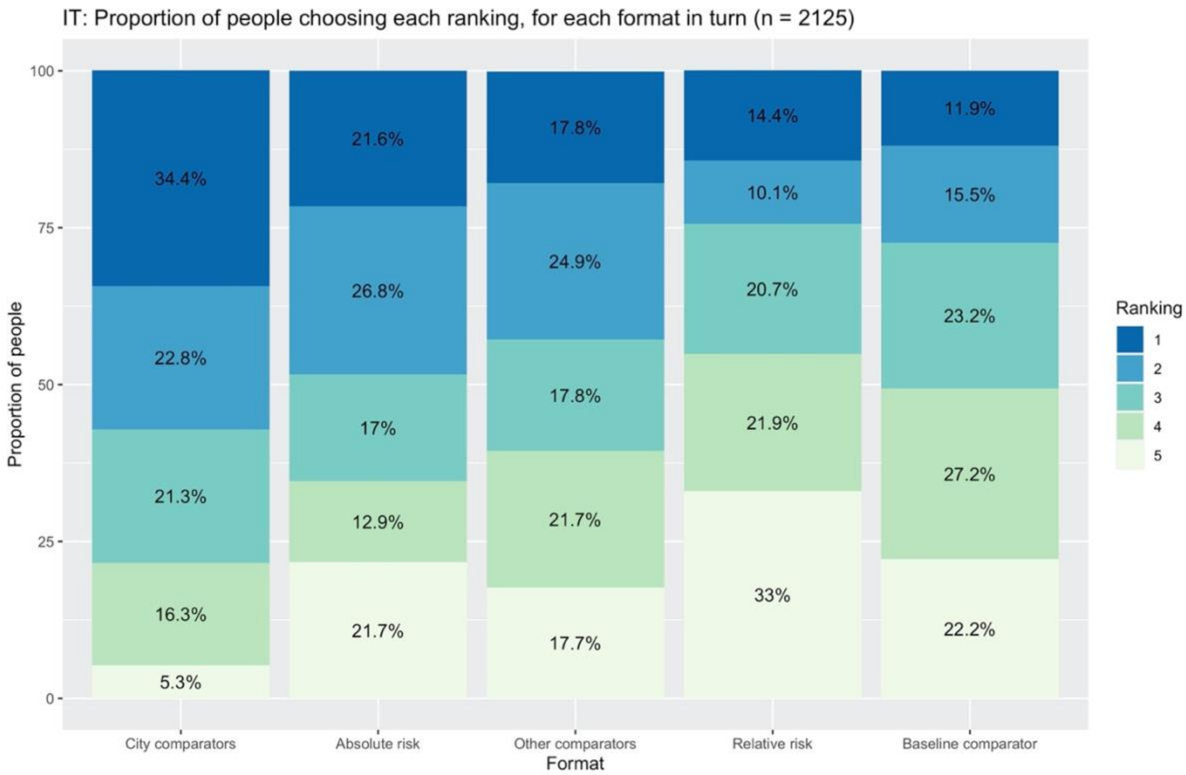


Figure 13: Participants' ranking of the different formats in the experiment when shown all of them at once. In Switzerland, a survey error meant that all but the last 350 participants were only shown four of the formats (they did not see the relative risk format), hence the Swiss plot detailed above contains just four formats. The final 350 Swiss participants who did see the relative risk format ranked it last in their favourites. Only rankings of participants who interacted with the drag-and-drop answer mechanism are included.

4.4. Conclusions

This experiment, especially the comments from participants seeing all the potential formats, highlights the difficulty of the balance between simple and clear information and enough information to help people interpret a number for themselves (i.e. without telling them what you think their interpretation 'should' be). It is possible that many of the participants, when asked which graphic they 'preferred', were not putting themselves in the position of trying to assess a hazard in real life and hence were prioritising simplicity rather than informativeness. When designing informative communications designed to support decision-making, we quite often find that members of the public are not used to authorities supporting them to make a decision themselves, and are more used to being 'told what to do' (something some people refer to as 'infantilisation').

It also highlights how much difference can be made to peoples' perception of the absolute risk by adding contextual information – such as a baseline risk, relative risk, or other comparator risks. It is interesting how the risk ladders we tested appeared to adjust people's relative perception of the risk i.e. they shifted people's perception upwards, compared with their perception of the absolute risk. In Switzerland however, this still resulted in lower overall perceptions of the same risk than in Italy or California. By contrast, giving the baseline risk for context appeared to have a slightly different effect – bringing the Swiss risk perception closer to the same absolute level as in Italy (this format was not tested in California). It might be tempting to think that this is because the Swiss participants had a differing baseline expectation from the Italians because they assumed a baseline from their own country, where it would be lower than in most parts of Italy. However, this wouldn't explain the different perception level in Switzerland compared with Italy when participants saw the 'other risks' risk ladder, which was identical between the two countries and thus by the logic above should have anchored their perception at the same absolute levels in the same way that the graphic showing the absolute baseline was designed to.

It is also interesting that the two risk ladder designs had different effects on participants' perception of the risks, across all three countries, possibly due to their prior beliefs about the risks used as comparators. Our elicitation of prior beliefs about the risk posed by each comparator suggests that the perception of the seismic risks was higher than that of the 'other' risks.

None of the formats had the effect found by Savadori and colleagues (Savadori *et al.*, 2022), where a risk ladder helped emphasise the differences between hazard levels by decreasing the perceived risk of low hazards (compared to a condition with no comparators) and increasing the perceived risk of high hazards (compared to a condition with no comparators). This is, however, in line with the findings of studies in other contexts, where comparator risks tend to shift general risk perceptions up and down, without changing the relative differences between hazard levels (e.g. (Keller and Siegrist, 2009; Pighin *et al.*, 2013)).

Looking at the preparatory actions that people reported that they would be likely to take, they very much echo the risk perception of each country. In Italy, the baseline comparison format (where the baseline hazard was said to be 0.003%) appears to result in a larger number of people taking no action at all in the face of a 6.7% risk of a damaging earthquake than other formats do. This, then, seems not to be giving people in Italy an appropriate impression of the hazard. The absolute plus relative risk format seemed not to help participants in Italy distinguish between the risk levels above baseline, and it was also ranked poorly by Italian participants (332 Swiss participants were also shown all five formats and ranked the relative risk format bottom).

Out of all the comparator formats, the risk ladder with comparator cities' seismic hazard marked on it is probably the one that enhances people's risk discrimination the most and gives them the most useful context. It was also consistently ranked highly overall. The choice of the comparator cities, though, needs to be careful (and information from the risk perception survey in this experiment may help guide choice of familiar cities of different hazard levels that could be used).

However, the responses of many participants in the free text suggest that a risk ladder (particularly the one with seismic city comparators) is a complex format and makes the information intimidating to some, despite it being ranked highly overall. It could therefore be argued that simply presenting the information in the percentage together with the geographical frequency format might be simpler to understand and result in a similar enhancement of the risk perception.

Future research could present participants with those two formats to compare effects directly in the same experiment, but out of the two, the combined percentage and geographical frequency format would seem the 'safest' as it clearly helped participants to distinguish between hazard levels of all magnitudes in Experiment 1 whilst the combined absolute (percentage) and relative risks format appeared to make the higher risks harder to distinguish, for Italian participants.

5. Experiment 3: Evaluating the best way to incorporate maps into a forecast communication

5.1. Introduction

Hazard maps are a common means of communicating long term seismic hazard (Marti, Stauffacher and Wiemer, 2019), but whether they can also be used for communicating dynamic hazard as part of operational earthquake forecasting has not yet been investigated. During our interviews and focus groups with members of the Italian public and other stakeholders (such as journalists and members of civil protection), maps were often requested as part of a mocked-up OEF dashboard, and people felt that maps were familiar and useful. The public-facing OEF system in New Zealand uses maps, with isolines and colours to delineate areas with different probabilities (see Figure 14).

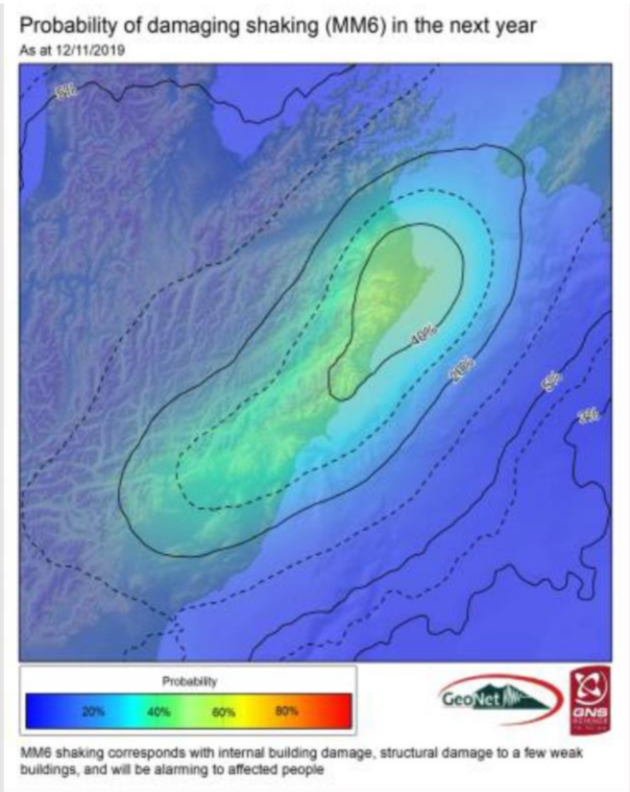


Figure 14: A probability distribution displayed in a map as part of New Zealand's public OEF display (https://www.geonet.org.nz/earthquake/forecast/kaikoura)

Maps allow the visualization of a hazard level (even multiple levels for different hazards) across an entire region, which could potentially be useful for planning (Carpignano *et al.*, 2009), or calls to action to trigger behaviour change (Dallo, 2022). However, work on the communication of natural hazard probabilities has already identified problems with the interpretation of such geographic representations, for example in maps showing volcano hazard probabilities (Thompson Clive *et al.*, 2021) and the 'cone of uncertainty' for hurricane track probabilities (Ruginski *et al.*, 2016).

A map that shows the probabilities of a seismic event affecting not just a single location, but surrounding locations as well is providing additional, contextual information to a forecast probability that might affect how the probability is interpreted by an audience interested mainly in information about their specific location.

We therefore wanted to test whether people's perception of a hazard level at a given location was affected by what else was being represented on a map of that location. In essence, we wanted to know whether people's interpretation of the hazard level at their own location would be influenced by how high it was 'relative to others'.

There could be two ways in which people's hazard perception might be affected by the other probabilities shown on a map.

1) Maps are commonly used to represent the effects of an earthquake that has already happened – the different intensities experienced, fanning out from an epicentre. We wondered whether this common usage would influence people's interpretation when maps with isolines were used for something very different – a forecast showing likelihoods of shaking events of a given intensity. During qualitative work examining possible OEF communication designs, although maps proved popular at first glance with audiences as they were a familiar format, we did find that some people naturally interpreted them as if they represented intensities rather than probabilities, as this was what they were used to. In a way this is analogous to people's interpretation of the hurricane cone of uncertainty representing the actual size of the storm (growing over time), which 80% of participants in one study did (Ruginski *et al.*, 2016; Padilla, Ruginski and Creem-Regehr, 2017). It is an interpretation of an illustration of probabilities (encompassing a range of possible scenarios) being interpreted as a concrete prediction of a single scenario.

   It could therefore be that some people's interpretation of the probabilities shown on a forecast map would be influenced specifically by the perception of their being an 'epicentre' formed by isolines: they might interpret the map directly analogously to a map showing intensities of a single event/scenario, rather than probabilities of a range of future ones. That would mean that those who saw their town shown apparently at an 'epicentre' (i.e. at the highest hazard level represented on the map) would perceive a higher risk than when they were shown their town at the *same absolute probability* of an earthquake, but no longer apparently the highest level on the map (i.e. no longer apparently at an 'epicentre').

2) An alternative way in which a map could affect perception would be simpler: that the mere presence of other hazard levels being represented would provide context for people to compare the hazard at their location with that of others (regardless of the shapes of the isolines). This would mean that if their town was represented as at the highest hazard level shown on the map, people might perceive the risk as higher than if it were not at the highest hazard level shown on the map.

There are also other pieces of contextual information given when a map is used that could affect audiences' assessments of the probabilities. An isoline map also requires a key, in order to help people identify the numerical hazard level represented at their location. In addition to the effects of the map design then, the design of the key can also give people an idea of relative hazard levels, by virtue of the chosen highest hazard level represented on the key and the position of the hazard in question relative to that highest point.

If any of these effects of relative context were manifested, then this could lead to bias in people's assessment of their risk, as an earthquake forecast should help people respond to the absolute hazard for their area, regardless of the hazard that others are experiencing (which will not affect their own area) or the possibility of higher hazards as represented by the key.

In this experiment, therefore, we aim only to identify whether people are responding to the absolute hazard level or the relative hazard level by manipulating the position of the isolines on the map and the nature of the key for the map. We also want to test whether a geographical map that does not have hazards represented on it as isolines might satisfy the audience's demand for geographical context without giving them unhelpful hazard context to the stated probability.

5.1. Methods

Participants were randomised to one of 4 arms, and each participant saw two stimuli within their respective arm in a randomised order. Hazard levels used in the experiment were chosen in order to allow us to use a linear scale on the key, and to illustrate enough absolute difference between the two hazard levels to be able to expect to detect a difference in participants' risk perceptions if they were viewing the absolute risks as percentages, but also to allow us to construct maps where a simple move of the isoline pattern would allow the same hazard level to be represented as relatively high and relatively intermediate in level compared with the surrounding areas. See Figure 15 for the stimuli used in each arm, in each country.

Arm 1 – Hazard levels 22% and 44% ('lower' and 'higher'), with cues from a key and from the map isolines, where isolines on both stimuli show the town in question as the highest hazard level ('no relative difference')

Arm 2 – Hazard levels 22% and 44% ('lower' and 'higher'), with cues from a key only. The map is geographic only, with no isolines ('isoline map cue absent').

Arm 3 - Hazard levels 22% and 44% ('lower' and 'higher'), with cues from the map isolines only (the hazard levels are shown numerically). The isolines on both stimuli show the town in question as the highest hazard level ('no relative difference').

Arm 4 – Hazard levels both 22% (both 'lower'), with cues from a key and from the map, where the isolines on one stimulus show the town in question as the highest hazard level, and the other shows it to be an intermediate level relative to surrounding areas ('relative difference').

The colours of the different hazard levels (purple) were chosen so as not to be easily confused with colours commonly used to represent geological/geographical features (green, blue, brown), not to be confused with alert levels (red, yellow, green), and to be distinguishable to those with a range of colour vision variations. Dark colours were used for higher hazard levels, as has been shown to be intuitive (Bostrom, Anselin and Farris, 2008).

We collected three main variables in California, and two in Switzerland and Italy:

1) **Risk perception**

Imagine that you visited a website that showed you this graphic indicating the chance of a damaging earthquake happening in your town (indicated with the grey marker) within the next 7 days. How would you classify that risk in your mind?

Move the slider below to indicate it: 'Very low risk' – 'Very high risk'

2) **Subjective comprehension** (measured once only per arm – for the 22% hazard level version of arms 1-3, and in the version of arm 4 that shows the hazard of the location at an intermediate level relative to surrounding areas), measured via two questions:

How much effort do you feel you had to put into understanding the graphic?

Move the slider below to indicate it: 'A lot of effort' - 'Very little effort'

How clear and easy do you find the graphic to understand?

Move the slider below to indicate it: 'Not at all clear' – 'Very clear'

3) **Worry** (measured only once per arm – for the 22% hazard level version of arms 1-3, and in the version of arm 4 that shows the hazard of the location at an intermediate level relative to surrounding areas) – only collected in California:

How worried would you be about the chance of a damaging earthquake happening in your town during the forecast period (6th-13th July) if you were shown this graphic?

Move the slider below to indicate it: 'Not at all worried' – 'Very worried'

We additionally collected potential covariates including education, objective numeracy, optimism/pessimism, fatalism and earthquake experience.

We pre-registered three formal hypotheses at https://osf.io/5kf9r/:

1) Participants seeing two stimuli representing a lower hazard and a higher hazard (i.e. arms 1, 2 & 3) will perceive their risk to be higher in the higher hazard situation.

2) Participants getting a cue from a scale as well as the map (arm 1) will perceive a greater difference in riskiness between the low and high hazard stimuli than those only being given a numerical hazard level next to the map (arm 3), and those only being given the cue from the scale with no hazard level isolines on their map (arm 2). We have no hypothesis regarding the difference between arms 2 and 3.

3) Participants being shown the same hazard level twice, but with different isoline patterns on the map (arm 4) will perceive the risk to be higher in the situation where the map shows the location of interest as the highest hazard level relative to other locations on the map, compared to the situation where the map shows the location of interest to be an intermediate hazard level relative to surrounding areas.

Figure 15: Stimuli used in Experiment 3 in a) California, b) Switzerland and c) Italy

5.1. Results

Risk perception

Figure 16 shows participants' ratings of how risky the first and second hazard level that they were shown felt to them. The first hazard level they were all shown was 22%. In arms 1-3 the second hazard level they were shown was 44%, and in arm 4 it was 22% again.

The results uphold our first hypothesis: participants seeing two stimuli representing a lower hazard and a higher hazard (i.e. arms 1, 2 & 3) perceived their risk to be higher in the higher hazard situation (see left hand side of Figure 16, which shows these within-subjects comparisons).

They also partially uphold our third hypothesis, about participants in arm 4. In this arm, participants were shown the same hazard level twice, but with different isoline patterns on the map. The first version of arm 4 they saw (arm 4a) showed the forecast location to be at intermediate hazard relative to other locations on the map. The second version of arm 4 they saw (arm 4b) showed the forecast location to be at the highest hazard level relative to other locations on the map. In Switzerland, participants rated arm 4b, where the isoline pattern showed the forecast location to be at the highest hazard level compared to other locations on the map, as slightly higher risk than arm 4a where the isoline pattern showed the forecast location to be an intermediate hazard level relative to surrounding areas (even though both were depicting the same hazard level – 22%). However, in California and Italy, participants perceived no significant difference between these two versions of arm 4 (see left hand side of Figure 16, which shows these within-subjects comparisons).

It is possible that the slightly higher risk perception of Arm 4b in Switzerland compared with that of Arm 4a is a statistical artifact, given that Arm 4b is identical to Arm 1a, and yet Arm 4b is also rated as slightly higher risk than Arm 1a in Switzerland. It could also be that at least some participants in Switzerland may be using the map to assess their own hazard level, rather than using the absolute hazard value displayed. Further research into this phenomenon could help confirm either way.

In partial support of our second hypothesis, there were some slight differences in sensitivity to changes in hazard level in the other three arms; looking at the within-subjects comparison graphs on the left-hand side of Figure 16, people seeing arm 3, which had no graphical key ('isolines only') and only showed percentage representations of the hazard values, seem slightly less sensitive to the difference between the 22% and 44% hazard levels than those who saw arms 1 or 2 (which both displayed a graphical key). This is not surprising: in Experiment 1 we also saw the effect of a bar graphic to accompany a percentage decreasing the perception of the 22% hazard, but decreasing perception of the 44% hazard to a far lesser degree i.e. the presence of the bar graphic increased the sensitivity of the difference between these hazard levels. The problem with a graphical representation is that it does not aid discrimination at very low probabilities, when the absolute difference is also very small.

What is perhaps more enlightening (and something that went against our hypotheses) however, is the lack of difference between arms 1 and 2: adding isolines to the map makes no difference to people's perception of (or sensitivity to) the hazard levels. The evidence from the results as a whole is that it seems likely that people are generally not using the isolines on the map in their assessment of the hazard.

Once again, we see lower overall perception of the risks in Switzerland, despite the same relative patterns emerging. To investigate whether this related to the amount of experience participants in each country had with earthquakes, we looked at those participants who saw the first stimulus (22%) in arm 2 of Experiment 3 (which showed just the key alongside a geographical map with no isolines). We explored whether risk perception of the forecast shown in this graphic varied depending on experience, and whether this relationship varied by country. We found a significant, positive relationship between experience with earthquakes and perception of the riskiness of the forecast ($F(1733) = 39.854$, $p<.001$, $b = 2.92$, $R2 = 0.02$). The strength of this relationship was consistent between countries; those with more experience with earthquakes tended to perceive the risk of the forecast as being higher, with Californian and Italian participants having more experience than Swiss participants, and thus perceiving the forecast to be higher risk.

a)

b)

c)

Figure 16: Participants' ratings of the perceived risk presented by the different graphical formats in a) California (n=2290), b) Switzerland (n=2383) and c) Italy (n=2326). On the left, within-subjects comparison of the first and second stimulus of their format participants were shown, faceted by experimental arm; on the right, between-subjects comparison of the format participants were shown, faceted by stimulus (a or b), to look at differences between formats [N.B. the second stimulus showed a 44% hazard level in arms 1-3, but 22% in arm 4]. Individual data points are plotted as jittered points, with means shown as dots with confidence intervals (+/- 95% CIs). Asterisks denote significant pairwise differences between groups, adjusted using the Benjamini-Hochberg procedure (p <.05; ** p <.01, *** p <.001).*

Subjective comprehension

Figure 18 shows how participants rated the different formats in terms of clarity and ease of understanding, and how effortful they felt each one was to understand. This reveals remarkable similarity across all three countries, and between formats. Given the different amounts of information being carried in the formats, especially between those with and without isolines, this is a surprising finding, and again suggests that the audience were probably not using information from the isoline map.

Worry

The results for participants' ratings for worry (measured at the 22% hazard level in California only) closely mirror those for their risk rating on that hazard. See Figure 17. This is why the measure was dropped for the surveys in Switzerland and Italy, which had become longer by the addition of another hazard level in Experiment 2.

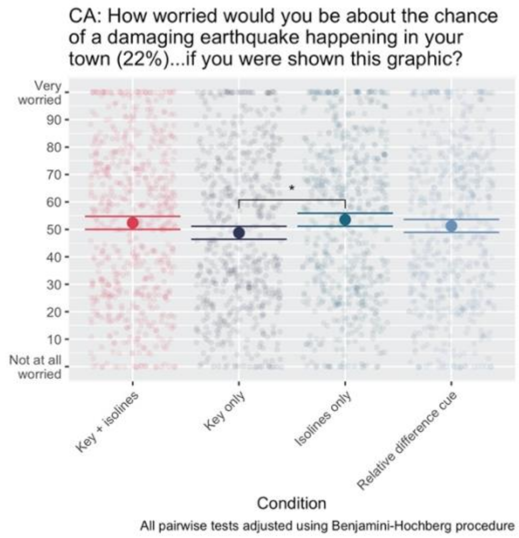


Figure 17: Participants' ratings of their worry at seeing a 22% hazard represented in each format in California. Individual data points are plotted as jittered points, with means shown as dots with confidence intervals (+/- 95% CIs). Asterisks denote significant pairwise differences between formats, adjusted using the Benjamini-Hochberg procedure (p <.05; ** p <.01, *** p <.001).*

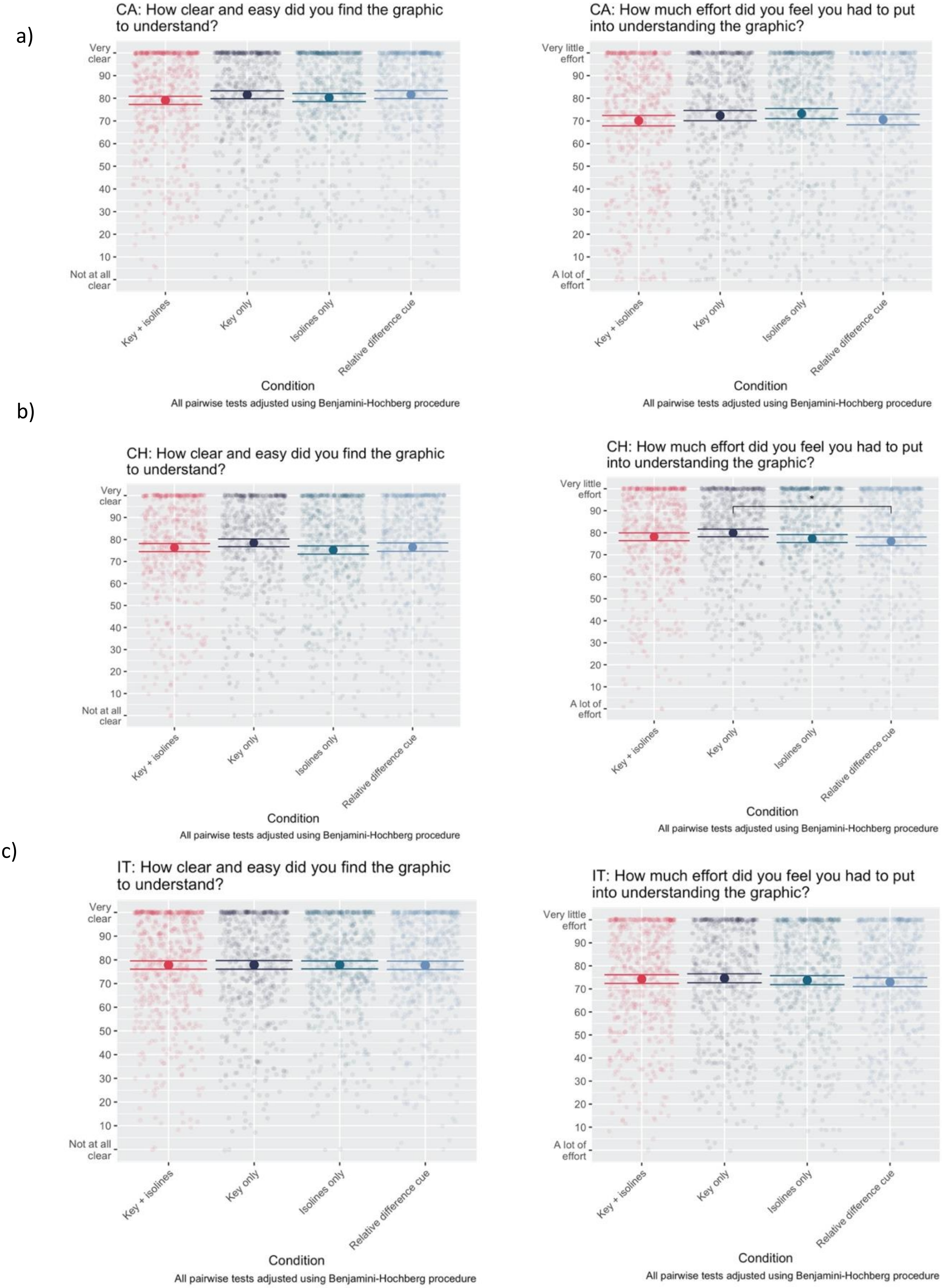


Figure 18: Participants' ratings of the clarity and effort required to understand each of the four formats in a) California (n=2290), b) Switzerland (n=2383) and c) Italy (n=2326). Individual data points are plotted as jittered points, with means shown as dots with confidence intervals (+/- 95% CIs). Asterisks denote significant pairwise differences between formats, adjusted using the Benjamini-Hochberg procedure (p <.05; ** p <.01, *** p <.001).*

5.2. Conclusions

Trying to represent forecast hazard levels in the style of isolines on a map, as is currently often attempted in operational earthquake forecasts such as that used in New Zealand, might be useful for an 'at a glance' summary for a policy-maker with a broad geographical area to consider. However, audiences who only need to know the forecast for their local area appeared not to use the isoline map information in our experiment – instead apparently taking their information from text or graphics stating the probability in their area.

However, the presence of even a simple, geographical map bearing no seismic information at all on the forecast may be perceived as useful to people – being a familiar format and allowing them to visualise the area over which the forecast is valid. We saw no significant difference in perception of the hazard levels we tested between those provided with the information about the probability accompanied by an isoline map ('key and isolines') or a normal geographical map ('key only').

Comparing the perception of the risk of 22% and 44% in California with the perception of the formats in Experiment 1 (see Figure 19), we can see that all the formats in this experiment lay just below the frequency formats in their risk perception (and above the plain absolute percentage format) for the 22% hazard level, and just above them all for the 44% hazard level. It is interesting to compare the 'key only' and the 'percentage plus bar graphic' which are similar in their design (using a graphical marker along a bar to indicate the probability). In Experiment 1 it was shown that the bar graphic is only really useful for aiding discrimination of these higher (>1%) probabilities. The higher rating of the risk when shown the bar key in Experiment 3 than the bar graphic in Experiment 1 could be due to the 'zoomed in' nature of the bar key in Experiment 3, meaning the marker's position on the bar was further to the right, or it could be a function of having the geographical map alongside it, which could make the probabilities (and their effects) seem more psychologically solid.

Further research could compare the effects of adding a simple geographical map alongside the different formats tested in Experiment 1, which we suggest may be the most practical and popular format for presenting OEF to the public.

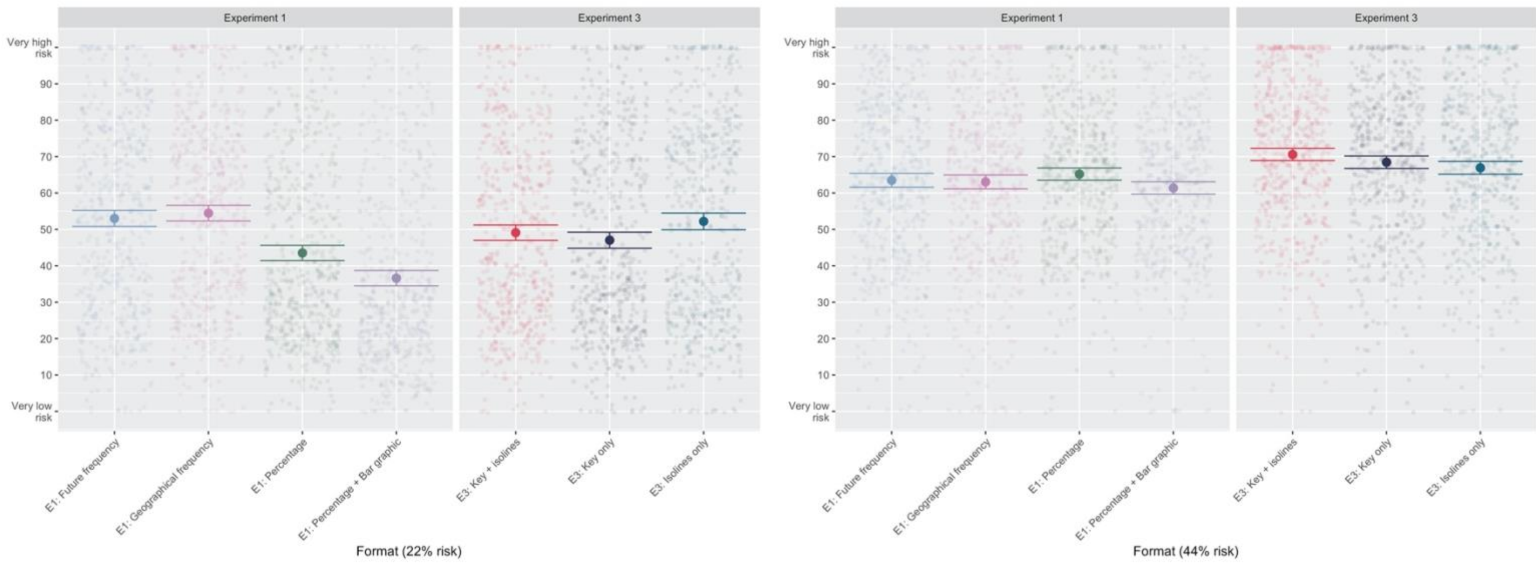


Figure 19: The ratings of Californians (n = 2290) to the 22% (left) and 44% (right) hazard levels in different formats across Experiments 1 and 3, plotted together for comparison. Individual data points are plotted as jittered points, with means shown as dots with confidence intervals (+/- 95% CIs).

6. Pulling it all together: conclusions

The challenges

Communicating information in such a way as to support someone else's independent decision-making, rather than simply trying to convey a 'take home message', presents many challenges. This is particularly true when the audience is broad and varied in terms of its knowledge about the topic, and the information needed for them to make such a decision has multiple aspects.

During our qualitative work with members of the public around the communication of seismic forecasts, we identified several challenges:

Some people did not understand the scientific basis for seismic forecasts, thinking that they were more like storm forecasts and hence either expecting a precise prediction of when and where an earthquake was expected to occur, or - realising that such a prediction was not possible - mistrusting any forecast information at all. This lack of background meant that the audience expected certain types of information that were not necessarily suitable for an OEF.

One expectation that many of the audience had, was that the information would be simplified to a warning or reassurance message. The degree to which information should be interpreted and translated into a message for an audience is one tension point. Setting the audience's expectations for the information they are about to receive and what they need to do with it (i.e. they need to make their own decision, rather than follow instructions) is clearly important.

Another clear tension point is the trade-off between amount of information and how understandable it is. Some members of the public wanted the minimum amount of information (such as just a single, absolute risk expressed as a percentage), whilst others wanted further information to help them interpret that information. This tension could to some extent be improved by setting expectations about the information and why it is being provided, but also it is clear that some further contextual information needs to be provided as an optional extra (clearly signposted) for those who want it.

One of the challenges about the level of information that needs to be included in an OEF is the need to try to communicate the impact as well as the probability of the earthquake, and the geographical area and time period over which the forecast is valid.

Formats that are familiar to an audience are much more easily parsed and understood, which is probably why maps are so often requested in our qualitative research. However, using a familiar format (such as a map) in an unfamiliar way (such as to represent probabilities rather than impacts) can lead to misunderstandings, although in this case it appears that people generally didn't use the map – probably because the information they were looking for was available to them in a simpler format elsewhere in the graphic.

However, familiarity can be a useful way to reduce the feelings of information overload, and so choosing one format in which to communicate seismic forecast information and repeatedly using that same format, on a daily basis and across different media outlets, could help audiences become used to how to interpret it (as well as becoming familiar with the baseline forecast and hence potentially more sensitive to changes from baseline). This means that the format chosen has to be functional and useful across many orders of magnitude of probability and whether the information is likely to result in alarm and action (such as in an aftershock sequence), or not (such as in a quiescent period).

Some overall findings

In these three experiments, we concentrated on specific aspects of the communication of seismic forecast information: the effects of different numerical formats on perception of the risk, the effects of different contextual information on perception of the risk, and the effects and interpretation of using maps to try to communicate probabilities of an earthquake. To do so, we did not attempt to communicate the impacts of the forecast earthquake (using 'damaging' as a descriptor, which we hoped would minimise variation in interpretation of the potential impact).

We found remarkably consistent results across three countries and languages, which provides reassurance that general principles established during work undertaken in one country is likely to apply in another. Having said that, perceptions of (the same) seismic risks, and hence the actions that people said they might take in response to them, were in general lower in Switzerland. This seems to be in part because people with personal or close experience of earthquakes have a higher perception of the risk that they present and there may be fewer people in Switzerland with this experience. However, many more factors are clearly at work. Further analysis of the data should shed more light on this.

How can we tell what is the 'best' format?

When trying to inform people's decision-making but not to change their behaviour in one particular direction (i.e. not trying to give them a 'call to action'), it is much more difficult to define and measure 'success'. What is the 'right' level of perception of a risk? Several authors have worked on criteria of success for such information provision (e.g. (Weinstein and Sandman, 1993; Weinstein, 1999; Michie *et al.*, 2002). Measures of subjective feelings of 'informedness' and decision satisfaction can be used to compare formats, and of course the ability to be able to tell the difference between two hazard levels is important. Measures of what behaviours people say they would take as a result of the information, however, can also be an important measure of someone's absolute perception of a risk. Although such answers in surveys are purely hypothetical, and also reflect someone's own personal values and vulnerabilities (e.g. if they live in a house that they think very prone to collapse they may legitimately have a much lower threshold at which they would consider evacuating than if they live in a house with high seismic resistance), how a person says they would behave can give a sense of whether someone is perceiving a risk in a similar way to how an expert would interpret it. It can give us a sense of whether a format that gives a 'higher' or 'lower' relative perception of the same risk is more appropriate.

Which numerical formats to use?

As was anticipated from previous research, we found in Experiment 1 that expressing a risk in the format of an expected frequency of the event (if there were x opportunities for the event to occur, how many times we would expect it to actually happen) made the risks seem significantly higher than expressing it as a simple percentage, until the highest probability level we tested (44%). The frequency formats were deemed harder to understand, which was also anticipated as they involve a lot more words and engage the imagination (which may explain their effects in raising perception of the risk), causing cognitive effort.

Expected frequency formats are commonly used in the communication of health outcomes where it is phrased as imagining multiple patients with exactly the same risk factors undergoing the same treatment or event, some having one outcome and others not. For a single or rare event such as an earthquake, the expected frequency format is not quite so easy to construct as it really involves imagining multiple potential ways in which the future might play out. However, for the purposes of a forecast where the risk applies only to a town or local area, we constructed a phrasing that instead mirrored the medical phrasing, with 'patients' replaced by 'towns'. This format proved to be the one that elicited the highest perception of the risk of each probability. Combining it with a percentage, so that participants saw both the percentage and the frequency interpretation of it, was deemed more easily understood than just the frequency phrasing alone, and the perception of the risk of this combined format was intermediate between the individual frequency and percentage formats alone.

This combination of frequency and percentage might be a case where 'more is more', and it seems that this combination gives a clear and understandable view of the probabilities, aiding discrimination of very low probabilities (where percentages with several decimal places are difficult to interpret).

Adding a graphic to help illustrate a probability is often considered an important part of risk communication. At higher probabilities, we found in Experiment 1 that presenting a graphic alongside a percentage showing the absolute risk significantly decreased risk perception (a 44% risk looked lower when it was illustrated by a graphic than when it wasn't). At low probabilities (the most common situation), it was impossible to illustrate such a low percentage graphically, and zooming in would defeat the object of the exercise, as it would change the cue that such a graphic gives (as was illustrated in Experiment 3 where, for convenience, we used a cut-off 0-50% scale and – as Figure 19 shows – the risk perception of participants was then higher than those in Experiment 1 who had been shown the graphic bar that was scaled from 0-100%).

Our recommended format for the numerical information is therefore:

"*With current levels of seismic activity the chance of an earthquake of [insert magnitude or intensity information here] happening in [insert location information here] between [insert dates here] is: x%.*

Imagine 100,000 places with exactly the same chance of an earthquake as [insert location information here].

In the week of [insert dates here], with an x% chance, we would expect:
An earthquake of [insert magnitude or intensity information here] to happen in y of these places.
No earthquake of [insert magnitude or intensity information here] to happen in z of these places."

What to display alongside the probability to help people interpret it?

Maps are popular ways to illustrate OEF probabilities. However, our Experiment 3 suggests that people interested in only the forecast at their location generally do not use a map with isolines (and there is very tentative evidence that it could lead to misinterpretations of a hazard level). We therefore advise that for public OEF communication, a map is used purely to illustrate the geographical area over which the forecast is calculated. This allows people to 'see' where the forecast is for (and potentially to move the map to navigate to other areas to read the forecast for those), whilst not being confused by the map trying to convey probabilistic information for other areas. It seems that probabilistic information for a single location is conveyed more clearly merely with words and numbers, as discussed above.

However, we also experimented with other forms of contextual information to help people interpret the risk of an earthquake in their area. The forms that we experimented with included several that have been previously recommended or tested, within seismic communication or other forms of risk communication: a comparison with the absolute baseline hazard, a relative risk (calculated from the comparison between the current forecast and the baseline risk for the area), or a comparison with other risks on a risk ladder: either other seismic hazards, or other non-seismic risks to life.

We expected the combination of an absolute and relative risk to give people the highest perception of the hazard level (above baseline), and it did – although the behavioural measures suggest that it was not resulting in extreme behavioural intentions. However, the results from Italy (although this was a between-subjects, not within-subjects experiment) suggest that people were not so well able to distinguish between different hazard levels above baseline in this combined absolute and relative risk format, which must give some caution to any conclusions about the format.

The effect of giving the absolute baseline hazard level for comparison was complex. In Italy, it appeared to give a lower risk perception for the higher hazard levels shown (3.9%; 6.7%) than the (high) perception given by the absolute risk alone; in Switzerland it appeared to give a higher risk perception for all hazard levels than the (low) perception given by the absolute risk alone. The cause of these differences between countries needs further investigation.

The effects of the two risk ladders, by contrast, were fairly consistent across hazard levels and between countries: they both increased people's perception of the risk, with the seismic city comparators increasing perception more than the other, non-seismic risk comparators (the differences made by the risk ladders being slightly larger in Switzerland and Italy than in California).

Subjectively, the seismic city comparator risk ladder was universally judged to be the most difficult to understand, which was anticipated because it includes the most complex and unfamiliar information. It also ranked low on satisfaction in providing information. However, for some people, according to the free text, it was appreciated as the most informative and suitable contextual addition. It also performed well in the ranking questions we asked in Switzerland and Italy, ranking first compared to the other formats in Italy and second in Switzerland.

From all these complex results, we can perhaps summarise in a table:

Format	Pros	Cons
Absolute risk alone, percentage format *"The chance… is x%"*	Rated subjectively clear Risk perception shows good response curve	Doesn't give any context Potentially too few people taking action at higher hazard levels
Absolute risk alone, geographical frequency format *"Imagine 100,000 places with exactly this chance… In the week, we'd expect…"*	Risk perception shows good response curve Potentially good number of people taking action at higher hazard levels	Rated subjectively unclear
Absolute risk alone, future frequency format *"Imagine 100,000 ways in which the week could turn out. With this chance… we'd expect…"*	Risk perception shows good response curve Potentially good number of people taking action at higher hazard levels	Rated subjectively unclear
Absolute risk alone, combined percentage & geographical frequency *"The chance… is x%. Imagine 100,000 places with exactly this chance… In the week, we'd expect…"*	Risk perception shows good response curve Potentially good number of people taking action at higher hazard levels Rated subjectively quite clear	
Absolute and relative risk *"The chance… is x%. This is y times higher than in a typical week"*	Rated subjectively quite clear Potentially good number of people taking action at higher hazard levels	In between-subjects experiment (Italy), decreased discrimination at hazard levels above baseline Subjectively ranked low
Absolute with baseline comparator *"The chance… is x%. In a typical week, the chance is y%"*		Gave mixed results in different countries: in Italy decreased discrimination of hazard levels Potentially too few people taking action at higher hazard levels
Absolute with seismic city comparators *"The chance… is x%. For context, the chance in a typical week in … is…* [graphic of different cities showing their position along a risk ladder]	Risk perception shows good response curve Potentially good number of people taking action at higher hazard levels Ranked subjectively high	Rated subjectively unclear
Absolute with other risk comparators *"The chance… is x%. For context, the chance of …* [graphic of different risks of death showing their position along a risk ladder]	Risk perception shows good response curve Potentially good number of people taking action at higher hazard levels	Rated subjectively unclear Ranked subjectively mid-low

Looking across all these findings, we recommend that the primary, front page information given in an OEF is simply the absolute current probability of a seismic event (of a particular size and in a particular area and in a particular timeframe), expressed as a percentage combined with a (geographic) frequency, alongside a map illustrating the area.

We then recommend that a clear link is given to ‘put this risk into context’ which would bring up the seismic city comparator risk ladder for those who want this further information, since out of those who did express a desire for it, the seismic city comparators seemed preferred over the ‘other risks’ comparators.

Practical considerations

Creating the mock-ups for the graphical formats tested in these experiments raised several practical issues.

As well as having to decide on and clearly communicate an appropriate size of area, timeframe and magnitude/intensity of earthquake to communicate the forecast over (all of which will affect the size of the probability being communicated, and hence the likely response to it), many seismic calculations need to be made.

How do we calculate the OEF for a given area, timeframe and earthquake magnitude/intensity? For practical purposes, people don’t just care about the probability of an epicentre occurring within that area, but also about felt earthquakes whose epicentres are outside of the area, so this needs to be calculated as well.

Then there is the need to source realistic comparator data. What is the baseline hazard for each area? Over what timeframe should that be calculated? How should it change if there has been seismic activity in the area in recent years? When giving seismic comparators, over what area is that calculated (each city may be much larger than the size of the area chosen for OEF calculations)? Which cities are the best-known comparators (and might some elicit a much higher level of concern because of recent memories of particular incidents)? We hope that this last question is addressed to some extent by the elicitation of risks within our Experiment 2, where we found that – with a few exceptions, such as L’Aquila within Italy – the relative hazard levels of most cities we tested were ‘correctly’ identified, on average, by participants worldwide.

From these experiments, then, our recommended communication format is illustrated in Figure 1, although it should be emphasised that this exact format has not yet itself been tested empirically. Around this central format for communicating the probabilities involved, the rest of the page design (communicating the other necessary information) also needs to be informed by our interview findings, which will be summarised in our final deliverable for the RISE project, the best practice guidelines.

References

Bostrom, A., Anselin, L. and Farris, J. (2008) 'Visualizing seismic risk and uncertainty: A review of related research', *Annals of the New York Academy of Sciences*, 1128, pp. 29–40. doi: 10.1196/annals.1399.005.

Camerer, C. F. and Kunreuther, H. (1989) 'Decision Processes for Low Probability Events: Policy Implications', *Journal of Policy Analysis and Management*, 8(4), pp. 565–592. doi: 10.3109/15569527.2012.667030.

Carpignano, A. *et al.* (2009) 'A methodological approach for the definition of multi-risk maps at regional level: First application', *Journal of Risk Research*, 12(3–4), pp. 513–534. doi: 10.1080/13669870903050269.

Covello, V. T. (1991) 'Risk comparisons and risk communication: Issues and problems in comparing health and environmental risks', in Kasperson, R. E. and Stallen, P. J. M. (eds) *Communicating Risks to the Public: International Perspectives*. Dordrecht: Springer Netherlands, pp. 79–124. doi: 10.1007/978-94-009-1952-5_6.

Covello, V. T., Sandman, P. M. and Slovic, P. (1988) *Risk Communication, Risk Statistics, and Risk Comparisons: A Manual for Plant Managers*. Washington, D.C.: Chemical Manufacturers Association.

Dallo, I. (2022) *Understanding the communication of event-related earthquake information in a multi-hazard context to improve society's resilience*. ETH Zurich.

Dryhurst, S. *et al.* (2020) *Communicating seismic forecasts*. Available at: http://static.seismo.ethz.ch/rise/deliverables/Deliverable_5.1.pdf.

Fischhoff, B. *et al.* (1978) 'How safe is safe enough? A psychometric study of attitudes towards technological risks and benefits', *Policy Sciences*, 9(2), pp. 127–152. doi: 10.1007/BF00143739.

Fischhoff, B. and MacGregor, D. (1983) 'Judged Lethality: How Much People Seem to Know Depends Upon How They Are Asked', *Risk Analysis*, 3(4), pp. 229–236. doi: 10.1111/j.1539-6924.1983.tb01391.x.

Freeman, A. L. J. *et al.* (2021) *Communicating personalized risks from COVID-19: Guidelines from an empirical study*, *Royal Society Open Science*. doi: 10.1098/rsos.201721.

Halpern, D. F., Blackman, S. and Salzman, B. (1989) 'Using statistical risk information to assess oral contraceptive safety', *Applied Cognitive Psychology*, 3(3), pp. 251–260. doi: 10.1002/acp.2350030305.

Keller, C. and Siegrist, M. (2009) 'Effect of risk communication formats on risk perception depending on numeracy', *Medical Decision Making*, 29(4), pp. 483–490. doi: 10.1177/0272989X09333122.

Keller, C., Siegrist, M. and Gutscher, H. (2006) 'The role of the affect and availability heuristics in risk communication', *Risk Analysis*, 26(3), pp. 631–639. doi: 10.1111/j.1539-6924.2006.00773.x.

Kunreuther, H., Novemsky, N. and Kahneman, D. (2001) 'Making Low Probabilities Useful', *Journal of Risk and Uncertainty*, 23(2), pp. 103–120. doi: 10.1023/A:1011111601406.

Lipkus, I. M. (2007) 'Numeric, verbal, and visual formats of conveying health risks: suggesting best practices and future recommendations', *Medical Decision Making*. Lipkus, Isaac M., Duke University Medical Center, Department of Psychiatry, 905 West Main Street, Box 34, Durham, NC, US, 27701: Sage Publications, 27(5), pp. 696–713. doi: 10.1177/0272989X07307271.

Marti, M., Stauffacher, M. and Wiemer, S. (2019) 'Difficulties in explaining complex issues with maps: Evaluating seismic hazard communication - The Swiss case', *Natural Hazards and Earth System Sciences*, 19(12), pp. 2677–2700. doi: 10.5194/nhess-19-2677-2019.

Michie *et al.* (2002) 'The multi-dimensional measure of informed choice: a validation study',

Patient Education and Counseling, 48(1), pp. 87–91. doi: 10.1016/S0738-3991(02)00089-7.
Padilla, L. M., Ruginski, I. T. and Creem-Regehr, S. H. (2017) ‘Effects of ensemble and summary displays on interpretations of geospatial uncertainty data’, *Cognitive Research: Principles and Implications*. Cognitive Research: Principles and Implications, 2(1), pp. 1–16. doi: 10.1186/s41235-017-0076-1.
Persoskie, A. and Downs, J. S. (2015) ‘Experimental Tests of Risk Ladders in the Elicitation of Perceived Likelihood’, *Journal of Behavioral Decision Making*, 28(5), pp. 424–436. doi: 10.1002/bdm.1858.
Peters, E., Hart, P. S. and Fraenkel, L. (2011) ‘Informing patients: The influence of numeracy, framing, and format of side effect information on risk perceptions’, *Medical Decision Making*, 31(3), pp. 432–436. doi: 10.1177/0272989X10391672.
Pighin, S. *et al.* (2013) ‘Using comparison scenarios to improve prenatal risk communication’, *Medical Decision Making*, 33(1), pp. 48–58. doi: 10.1177/0272989X12464433.
Ruginski, I. T. *et al.* (2016) ‘Non-expert interpretations of hurricane forecast uncertainty visualizations’, *Spatial Cognition and Computation*, 16(2), pp. 154–172. doi: 10.1080/13875868.2015.1137577.
Sandman, P. M., Weinstein, N. D. and Miller, P. (1994) ‘High Risk or Low: How Location on a “Risk Ladder” Affects Perceived Risk’, *Risk Analysis*, 14(1), pp. 35–45. doi: 10.1111/j.1539-6924.1994.tb00026.x.
Savadori, L. *et al.* (2022) ‘Communicating Seismic Risk Information: The Effect of Risk Comparisons on Risk Perception Sensitivity’, *Frontiers in Communication*, 7(March), pp. 1–11. doi: 10.3389/fcomm.2022.743172.
Siegrist, M. (1997) ‘Communicating low risk magnitudes: Incidence rates expressed as frequency versus rates expressed as probability’, *Risk Analysis*, 17(4), pp. 507–510. doi: 10.1111/j.1539-6924.1997.tb00891.x.
Siegrist, M., Orlow, P. and Keller, C. (2008) ‘The effect of graphical and numerical presentation of hypothetical prenatal diagnosis results on risk perception’, *Medical Decision Making*, 28(4), pp. 567–574. doi: 10.1177/0272989X08315237.
Slovic, P., Monahan, J. and MacGregor, D. G. (2000) ‘Violence risk assessment and risk communication: The effects of using actual cases, providing instructions, and employing probability vs. frequency formats’, *Law & Human Behavior*, 24(3), pp. 271–296. doi: 10.1023/A:1005595519944.
Thompson Clive, M. A. *et al.* (2021) ‘Volcanic hazard map visualisation affects cognition and crisis decision-making’, *International Journal of Disaster Risk Reduction*. Elsevier Ltd, 55, p. 102102. doi: 10.1016/j.ijdrr.2021.102102.
Weinstein, N. D. (1999) ‘What does it mean to understand a risk? Evaluating risk comprehension.’, *Journal of the National Cancer Institute. Monographs*, (25), pp. 15–20. doi: 10.1093/oxfordjournals.jncimonographs.a024192.
Weinstein, N. D. and Sandman, P. M. (1993) ‘Some criteria for evaluating risk messages’, *Risk Analysis*. Blackwell Publishing Ltd, 13(1), pp. 103–114. doi: 10.1111/j.1539-6924.1993.tb00733.x.
Wilson, R. and Crouch, E. A. C. (1987) ‘Risk assessment and comparisons: An introduction’, *Science*, 236(4799), pp. 267–270. doi: 10.1126/science.3563505.

Appendix

Table A1: Key demographic variables across surveys. Note that the answer options participants could choose from for the education question were adapted to the study location in question - the options for California are shown in this table.

		California		Switzerland		Italy	
		n	%	n	%	n	%
Gender	Male	1088	*47.5*	1130	*47.4*	1085	*46.6*
	Female	1195	*52.2*	1249	*52.4*	1241	*53.4*
	Other	7	*0.3*	4	*0.2*	0	*0*
Age	18-24	260	*11.4*	193	*8.1*	196	*8.4*
	25-34	434	*19*	389	*16.3*	303	*13*
	35-44	354	*15.5*	402	*16.9*	416	*17.9*
	45-54	408	*17.8*	435	*18.3*	534	*23*
	55-64	461	*20.1*	470	*19.7*	461	*19.8*
	65+	373	*16.3*	494	*20.7*	416	*17.9*
Education	No formal qualifications	35	*1.5*	12	*0.5*	119	*5.1*
	High School Diploma	620	*27.1*	632	*26.5*	1285	*55.2*
	Associate Degree or Certificate	405	*17.7*	644	*27*	26	*1.1*
	Bachelors Degree	659	*28.8*	351	*14.7*	262	*11.3*
	Masters Degree	336	*14.7*	263	*11*	494	*21.2*
	Doctoral Degree	112	*4.9*	49	*2.1*	89	*3.8*
	Other qualifications	62	*2.7*	350	*14.7*	26	*1.1*
	Prefer not to answer	52	*2.3*	78	*3.3*	20	*0.9*
	Missing	9	*0.4*	4	*0.2*	5	*0.2*
Numeracy score	1	267	*11.7*	122	*5.1*	148	*6.4*
	2	380	*16.6*	258	*10.8*	290	*12.5*
	3	454	*19.8*	425	*17.8*	510	*21.9*
	4	494	*21.6*	435	*18.3*	522	*22.4*
	5	333	*14.5*	470	*19.7*	383	*16.5*
	6	215	*9.4*	364	*15.3*	294	*12.6*
	7	73	*3.2*	144	*6*	95	*4.1*
	8	74	*3.2*	165	*6.9*	84	*3.6*