Accepted in Journal of Cognitive Psychology (Nov 2018).

This is the version of the article before publication.

Running head: PROBABILITY OF PRECIPITATION FORECAST

Not as gloomy as we thought

Reassessing how the public understands probability of precipitation forecasts

Original research submission

Word count: 7986

Author's note: The materials and data from both studies are available on the Open Science Framework: https://osf.io/76zes/?view_only=3f770f3a12634f128185b1ef7c8cb3e3.

Accepted in Journal of Cognitive Psychology (Nov 2018).

This is the version of the article before publication.

Abstract

Prior research asking people to interpret probability of precipitation (PoP) forecasts showed that many of them wrongfully believe that PoP forecasts are derived from a percentage of time, a percentage of a region or the strength of agreement among forecasters. We posit that the wording of PoP interpretation tasks matters, because it is associated with different metacognitive feelings used as cues in situations of uncertainty. We hypothesised that the fluency of the correct PoP interpretation is lower than the fluency of the incorrect interpretations and will, in turn, increase preference for the incorrect interpretations. We assessed the role of fluency in correctness perception (Study 1) and reassessed PoP interpretations with a more fluent correct interpretation (Study 2). Fluency perception was positively related with perception of correctness. Furthermore, participants selected the correct fluent interpretation more often than the correct disfluent one. We have drawn a more optimistic picture of people's PoP forecasts understanding than that shown before and have discussed the methodological and applied implications.

Key words: Probability of precipitation, weather forecast interpretation, fluency.

Accepted in Journal of Cognitive Psychology (Nov 2018).

This is the version of the article before publication.

Introduction

Probabilistic weather forecasts have been linked to wide economic benefits (Nadav-Greenberg & Joslyn, 2009; Society, 2008) and are sought by the public (Hanrahan & Sweeney, 2013; Morss, Demuth, & Lazo, 2008). However, the positive impact of probabilistic forecasts can only be fully achieved if people understand probabilities correctly. Past research has shown that people often misinterpret the set of events from which a probabilistic forecast is derived and, in turn, misunderstand the outcome that is being predicted. This is called the reference class problem.

In the present paper, we have assessed how people interpret probability of precipitation (PoP) and explore whether the nature of the task used to measure this interpretation affects participants' ability to identify the reference class of PoP.

What is the reference class of probability of precipitation forecasts?

Probabilistic weather forecasts are one of the most common pieces of probabilistic information that people are exposed to. PoP forecasts are among the top three concerns of weather forecast users (Lazo, Morss, & Demuth, 2009). The findings of a large nationwide survey conducted in the US (N = 1,520) showed that almost every adult in the US uses weather forecasts (96%) and that those adults are on average exposed to almost four weather forecasts every day (Lazo et al., 2009) which means that, every day, American adults are exposed to almost a billion weather forecasts.[1]

Most past research on PoP interpretation considered the reference class of a PoP for the day after to be "days that are like tomorrow", hereafter described as the "Days" interpretation (Gigerenzer, Hertwig, van den Broek, Fasolo, & Katsikopoulos, 2005;

[1] Number of adult US citizens based on a 2014 census multiplied by the proportion of American adults looking at weather forecasts, multiplied by the average number of daily forecasts on average $245\,300\,000 \times 0.964 \times 3.8 = 898\,582\,960$).

Hanrahan & Sweeney, 2013; Joslyn, Limor, & Nichols, 2008; Morss et al., 2008; Zabini, Grasso, Magno, Meneguzzo, & Gozzini, 2015). The Days interpretation has been used as the correct PoP interpretation in research conducted in a range of European countries such as Italy, Germany and the UK, as well as in the USA.

The Days interpretation suggests that PoP forecasts are derived from an ensemble of weather models, which is aligned with more formal definitions provided by national weather agencies (e.g., MetOffice, 2018). Given that the weather can evolve in many ways, weather forecasters generate many simulations, varying some of the parameters of the simulation slightly, such as wind strength. The output of the process is a set of simulations, known as an ensemble of forecasts, in which some feature rain but others do not. The probability of rain is mainly derived from the proportion of simulations of the weather where at least 1mm of precipitation occurs, anywhere in the forecasted area, at any time during the defined time period (MetOffice, 2018; National Weather Service, 2018b). If there is an agreement that "days like tomorrow" represents a correct reference class for PoP, there is less agreement regarding the precise process used to derive probability of precipitation; in fact this process may vary slightly across agencies (e.g., more or less adjustment based on meteorologists' experience, using the rain coverage or not). The way PoP forecasts are computed is not always made transparent and available to the public (nor to scientists) which may be because there is not a single PoP calculation method on which all weather forecast professionals and scientists agree (American Meteorological Society, 2008; De Elia & Laprise, 2005; Stewart et al., 2016).

A poor level of understanding of probability of precipitation forecasts (PoP)

Despite people's interest in and frequent encounters with PoP, more than half of the people surveyed and up to 80% provided an incorrect interpretation (Gigerenzer et al., 2005;

Joslyn et al., 2008; Joslyn, Nadav-Greenberg, & Nichols, 2009; Juanchich & Sirota, 2016; Morss et al., 2008; Murphy, Lichtenstein, Fischhoff, & Winkler, 1980; Sink, 1995; Zabini et al., 2015). The most common misinterpretation of PoP is to think that it refers to a proportion of time when it will rain (hereafter the "Time" interpretation). In the Time interpretation, the reference class of a PoP is a set of hours where precipitation occurred out of the total number of hours of the forecast. For example, in the case of a forecast for a 10% chance of rain over a 10-hour period, a Time interpretation means that it will rain for sure, but for only for 1 hour out of 10. The second most common error of interpretation of a PoP is to think that it refers to the proportion of a region that will receive rain (hereafter the "Region" interpretation). In the Region interpretation the reference class could be, for example, the proportion of square kilometres that received rain in the region of the forecast. For example, in the case of a forecast of a 10% chance of rain in a region measuring 1,000km^2, a Region interpretation would mean that it will rain for sure, but only over 100km^2. Note that in the PoP provided by the American National Weather Service, the Region interpretation can be correct, albeit in a single circumstance: if the probability of rain is 100% in an area. However, it is incorrect whenever the probability is lower than 100%. This is because the National Weather Service weighs the probability of rain by the forecast area (National Weather Service, 2018a). A final interpretation is the "Forecasters" interpretation, where the reference class represents a set of forecasters' opinions (Gigerenzer et al., 2005; Morss et al., 2008; Zabini et al., 2015). Therefore a 10% PoP would mean that 10% of the forecasters believe that it will rain. Most of the research is focused on the probability of rain, but findings are similar with other types of precipitation such as snow or hail (Juanchich & Sirota, 2016). A recent survey conducted on a large sample in the US (N = 1,337) showed that 3 participants out of 10 believed that PoP referred to the duration of a precipitation event (Time interpretation) and 2 out of 10

believed that it referred to the region receiving rain (Region interpretation; Juanchich & Sirota, 2016).

Misinterpretation of the reference class can lead people to erroneous probability perception, which, according to decision-making theories, might lead to ill-informed decisions (Morgenstern & von Neumann, 1943). People who wrongfully believe that a 30% probability of rain refers to a proportion of time are *certain* that it will actually rain tomorrow; they are only uncertain about when, which may result in suboptimal decisions. Many decisions based on weather forecasts may seem trivial and low risk (e.g., taking an umbrella, getting soaked), but some are actually very consequential. In fact, the financial success of some business activities directly depends on the weather (Lazo, Lawson, Larsen, & Waldman, 2011). Dutton suggests that one third of the US GDP (around $5.65 trillion in 2008), depends on the weather and hence can be sensitive to how it is forecasted (2002). For example, the agricultural industry may experience losses following important precipitation events and could mitigate those by adapting the timing of their harvest based on weather forecasts. However, if farmers misunderstand a 20% PoP as meaning that "it will rain for sure" they may incorrectly decide to postpone harvesting and lose part of their crop. Clearly, interpreting PoP correctly is critical for optimal decision-making and PoP misinterpretation could cause large economic losses.

The determinants of PoP interpretation: focus on the probabilistic format

Studies conducted to date have assumed that the ambiguity of the reference class has underpinned PoP misinterpretation. Studies aiming to improve PoP interpretation have essentially focused on using different formats and explanations to disambiguate the reference class (Gigerenzer et al., 2005; Joslyn et al., 2008; Juanchich & Sirota, 2016; Murphy et al., 1980). Yet, these manipulations have yielded only slight, if any, improvements in the level of

correct interpretation, leaving a lot of room for improvement. We propose to tackle the issue of poor understanding of PoP with another approach. We do not locate the source of the misinterpretation in the probability format, nor in the individual. We suggest, instead, that the measurement of the PoP interpretation itself could account for some variations in PoP interpretation performance.

Assessment methods of probabilistic weather forecast interpretation

In most of the previous research, the interpretation of weather forecasts was mainly assessed with open or multiple choice questions, often both measured within the same study (Gigerenzer et al., 2005; Juanchich & Sirota, 2016; Morss et al., 2008; Murphy et al., 1980). With open-ended questions, participants were simply asked to describe the "meaning" of a given forecast. When faced with such a task, most participants simply restated the probability in another format (e.g., writing "rain is likely" when given a 70% chance of rain) (Gigerenzer et al., 2005; Joslyn et al., 2009; Morss et al., 2008; Murphy et al., 1980; Peachey, Schultz, Morss, Roebber, & Wood, 2013). Only a minority of responses actually addressed the question of which particular event was uncertain (e.g., 23 out of 169 in Experiment 2; Joslyn et al., 2008).

In contrast, with multiple choice questions, participants are constrained to select one interpretation from a list (Gigerenzer et al., 2005; Juanchich & Sirota, 2016; Morss et al., 2008; Murphy et al., 1980). For example, in the study of Gigerenzer et al. (2005, p. 625) participants were prompted to select one of the three following Time, Region and Days interpretations for the forecast, "There is a 30% chance of rain tomorrow":

- *It will rain tomorrow for 30% of the time.*
- *It will rain tomorrow in 30% of the region.*
- *It will rain in 30% of the days like tomorrow.*

The multiple choice questions used in different studies varied in terms of probability magnitude of rain (e.g., 30%, 60%), number of options (between 2 and 5) and, most importantly, in the way the interpretations listed were phrased. For example, in Morss et al. (2008) and in Gigerenzer et al. (2005), the correct interpretation was worded as "it will rain on 60% of the days like tomorrow", whereas in Murphy et al. (1980) the correct interpretation was "the occurrence of precipitation at a particular point in the forecast area" (p. 697) and in Juanchich and Sirota (2015), the correct interpretation was "at least a minimum of rain will fall on 30% of the days like tomorrow" (p. 7). The variation in wording may seem inconsequential, but indirect evidence suggests that these wording variations led to changes in participants' responses. For example, Morss et al. (2008) provided two multiple choice questions to their participants with slightly different response choices[2]. The results consistently show a low rate of correct interpretation: only 29% of the participants selected the correct Days interpretation in the first multiple choice question, 19% selected it in the second and only 7% of the participants selected both correct interpretations. This difference was noted by the authors who hypothesised that it was caused by the difference in the wording of the choices.

A new approach to the investigation of weather forecast interpretation

In line with the possibility raised by Morss et al. (2008), we expected that the phrasing of the correct interpretation – and more precisely its fluency - could be responsible for the low adhesion rate to this interpretation (e.g., "at least a minimum amount of rain will

[2] Here is the exact wording of the choices provided in Morss et al. (2008, pp. 980-981). MCQ1: "It will rain tomorrow in 60% of the region", "It will rain tomorrow for 60% of the time", *"It will rain on 60% of the days like tomorrow"*, and "60% of weather forecasters believe that it will rain tomorrow." MCQ2: "It will likely rain over the entire forecast area tomorrow", "It will likely rain throughout the day somewhere in the forecast area tomorrow", "*It will likely rain at any one particular point in the forecast area tomorrow*", and "Weather forecasters are likely to believe that it will rain tomorrow." Both questions also featured two extra choices: "I don't know" and "Other (please explain)". The correct PoP interpretations are presented here in italics.

fall in 60% of the days like tomorrow"). Fluency is a metacognitive process that one experiences while processing information. It is defined as "the subjective experience of ease with which people process information" (Alter & Oppenheimer, 2009, p. 219). Fluency is not an objective indicator of how much effort people put into understanding a concept but rather an impression about their cognitive effort. The fluency one experiences when reading a sentence has several contributors, including the simplicity of the wording (e.g., smart vs. erudite; Oppenheimer, 2006) and how easy it is to imagine (e.g., a tree vs. an idea; Petrova & Cialdini, 2005). In the case of PoP interpretation, we hypothesised that a proportion of Time, Region or Forecasters is simpler and easier to imagine than a proportion of "days like tomorrow". We further hypothesised that this lower fluency was used by participants as a cue that the Days interpretation was not correct.

The fluency of a statement can indeed be used as a cue to its truthfulness and quality (Oppenheimer, 2006; Reber & Schwarz, 1999). For example, Oppenheimer (2006; Experiment 1) showed that a personal statement composed of many long words was considered to be poorer quality compared to a personal statement composed of simpler words (e.g. recognise vs. know). This effect was mediated by the perceived comprehensibility of the statement. Further, vacation destinations that were easier to picture were found to be more attractive than destinations that were difficult to imagine (Petrova & Cialdini, 2005). Building on these fluency findings, we hypothesised that the correct Days PoP interpretation was less fluent than the Region and Time interpretations and that this was deterring participants from selecting it as the correct answer.

Perceived fluency could also have an indirect effect on correctness perception that would be mediated by the perceived utility of the forecast. The Days interpretation could be perceived as less fluent, and thus less useful, and, in turn, less correct. Some evidence indeed

indicates that fluency might guide evaluation judgments. For example, stocks that have a fluent name were deemed to be more expensive than stocks with disfluent names (Alter & Oppenheimer, 2006; Borges, Goldstein, Ortmann, & Gigerenzer, 1999). Past work indeed suggests that the Time and Region interpretations could be judged more useful because the duration and location of rain are among the top three points of interest for people looking at weather forecasts (Lazo et al., 2009). It is also possible that these interpretations could be perceived as being more useful than the Days interpretation, because they suggest a 100% certainty of rain: the uncertainty only being about where and for how long it will rain.

If fluency predicts interpretation choice, participants in previous studies on PoP selected an option that did not fully reflect their conceptual interpretation of probability but rather a circumstantial preference driven by metacognitive feelings. The rate of PoP misinterpretation due to lack of conceptual understanding could therefore be lower than actually observed in previous studies.

Research overview

In Study 1, we tested whether the fluency – measured as comprehensibility and imaginability – of the Time, Region, Forecasters and Days PoP interpretations predicted the correctness of their perception. Furthermore, we suggested that the effect of fluency on correctness perception was mediated by the perception of the utility of the interpretation. Building on these findings, Study 2 aimed to reassess the ability of people to interpret PoP given an improved PoP interpretation task.

Open Science statement. The materials and data from both studies are available on the Open Science Framework: https://osf.io/76zes/?view_only=3f770f3a12634f128185b1ef7c8cb3e3.

Pilot study

Accepted in Journal of Cognitive Psychology (Nov 2018).

This is the version of the article before publication.

Development of a new and more fluent correct interpretation

We developed a correct interpretation that was designed to be conceptually similar but more fluent than the Days interpretation. The new interpretation was inspired by the MetOffice's definition of PoP: the current weather conditions are entered into a model and are used to extrapolate an ensemble of models of how the weather could evolve for the target period and area (MetOffice, 2018)[3]. The new interpretation was designed to be fully consistent with the Days interpretation which was used as the correct interpretation in past research (Gigerenzer et al., 2005; Hanrahan & Sweeney, 2013; Joslyn et al., 2008; Morss et al., 2008; Zabini et al., 2015). The Days interpretation can be understood as being slightly more generic than the Simulations, because "days like tomorrow" can refer to real past days like tomorrow, whereas the Simulations interpretation only refers to days that are simulated, which is the main source of information to compute PoP.

The new Simulations interpretation read: *"at least a minimum amount of rain will fall according to 30% of the ensemble of tomorrow's forecasts"*. This interpretation was judged to be more fluent than the Days interpretation in an informal pretest conducted with a few British native English speakers. This new interpretation is hereafter labelled the "Simulations" interpretation because it refers to an ensemble of computer simulations of how the weather will be the following day.

The Simulations interpretation is longer than all of the other interpretations (18 words): +3 words compared to the Time, Region and Forecasters interpretations and +2 words compared to the Days interpretation. Although there is no scientific evidence that the

[3] Here is the exact wording: "A weather forecast is an estimate of the future state of the atmosphere. It's created by observing the current state of the atmosphere and using a computer model to calculate how it may change over time. […] To estimate the uncertainty in the forecast we use what are known as Ensemble Forecasting. Here, we run our computer model many times from slightly different starting conditions."

length of a sentence is linked with its fluency (for a review, see Alter & Oppenheimer, 2009), this possibility is intuitively appealing. The length of the Simulations interpretation may have reduced the fluency of the option, but we expected that the increase in comprehension and ease of imaginability would nevertheless lead to an increase in fluency.

Assessing that the interpretations refer to different reference classes

To assess whether the Time, Region, Forecasters, Days and Simulations interpretations were perceived as referring to a different reference class, we conducted a short study (N = 121). Participants read the five interpretations of a PoP (presented in a random order to each participant) and could select one of five reference classes for each interpretation (*"things that were most likely used to compute the probability*"). The reference classes that participants could choose from, were number of minutes, square metres, forecasters, simulations and days like tomorrow. Participants were informed that they could select the same items twice or more. Results shown in Table 1 showed that most participants (>= 60%) understood that each probability interpretation entailed a specific reference class and selected the corresponding reference class (e.g., square metres for the Region interpretation; minutes for the Time interpretation…).

<Place Table 1 about here>

Table 1. Proportion of reference class selected (things that could be used to compute that probability) for the different PoP interpretations used in Experiment 1.

	Reference class				
	Minutes	Square	Forecasters	Days like	Simulations

Accepted in Journal of Cognitive Psychology (Nov 2018).

This is the version of the article before publication.

		metres		tomorrow	
Interpretation					
Time	**60%**	3%	6%	25%	7%
Region	7%	**63%**	5%	21%	5%
Forecasters	7%	3%	**78%**	12%	1%
Days	7%	4%	3%	**69%**	18%
Simulations	5%	2%	29%	5%	**60%**

Study 1

The aim of Study 1 was to assess whether the fluency of the PoP interpretations provided as possible answers guided participants' responses in the multiple choice questions. The study was designed to test the following hypotheses: (1) The fluency of the Days interpretation will be lower than the fluency of the three incorrect interpretations and lower than that of the correct newly developed interpretation. (2) The fluency of the Days interpretation will predict correctness perception and this effect will be mediated by the perceived utility of the interpretation. (3) The differences in fluency between the correct and incorrect interpretations will predict the difference in their perceived correctness and this effect will be mediated by the difference in their perceived utility.

Method

Participants. A total of 92 Americans from Amazon Mechanical Turk (AMT) took part in the survey. We determined the sample size following a rule of thumb based on past studies with similar designs. A post-hoc sensitivity analysis showed that our sample size

allowed us to detect a small effect size of $f = 0.11$ for a within-subjects ANOVA (assuming $\alpha = .05$, $1\text{-}\beta = .80$ and $r = 0.5$ between measurements).

Participants had a greater than 80% success hit rate (i.e., the measure of participants' performance specific to AMT) and fulfilled the involvement criteria to participate (i.e., to be an American native English-speaking adult). Amazon Mechanical Turk provides a valid pool of research participants (Buhrmester, Kwang, & Gosling, 2011; Goodman, Cryder, & Cheema, 2012; Paolacci, Chandler, & Ipeirotis, 2010). The participants' ages ranged from 18 to 63 ($M = 34.1$, $SD = 11.7$), after excluding the responses of two participants who were deemed to have mistakenly reported their ages (two and four years old). In the sample, 57% of the participants were female. Most participants had some experience of higher education, had a job and were white Caucasian. Table 2 provides an overview of the characteristics of the sample in Study 1 along with the characteristics of the American population.

Table 2. Characteristics of the samples of Study 1 and 2 compared to the general American population. Our participants are younger and more educated than the general American population.

Demographic characteristics	Study 1	Study 2	American population
Median age	30	30	38[a]
% of women	57%	57%	51%[b]
% of white ethnicity	78%	77%	73%[b]
% with (at least) a college degree	90%	81%	59%[c]
% working	69%	66%	63%[d]
N	92	114	322,311,308[e]

Accepted in Journal of Cognitive Psychology (Nov 2018).

This is the version of the article before publication.

Note: Data from the United States Census Bureau for 2016 (United States Census Bureau, 2018).
[a] From the 2016 Nation's Median Age graphic. The figure includes children.
[b] From the QuickFacts table: Population Estimates, July 1, 2017.
[c] Educational Attainment of the Population 18 Years and Over, by Age, Sex, Race, and Hispanic Origin: 2016.
[d] Employment Status, 2016 American Community Survey 1-Year Estimates (S2301). In civilian labor force, total, percent of population age 16 years+, 2012-2016.
[e] From the US and World Population Clock for 2016.

Materials and procedure.

In a web survey, participants provided a series of judgments for five PoP interpretations: Time, Region, Forecasters, Days and Simulations (see Table 3). The Time, Region, Forecasters and Days interpretations were similar to those used by Gigerenzer et al. (2005) and Morss et al. (2008). The Time, Region and Forecasters interpretations were incorrect, whereas the Days and Simulations were correct.

Table 3. Common interpretations of probability of precipitation (PoP) forecasts. The Time, Region and Forecasters interpretations are incorrect whereas the Days and Simulations are correct.

Label	Interpretation
Time	Tomorrow it will rain for 30% of the time
Region	Tomorrow it will rain in 30% of the region
Forecasters	It will rain tomorrow according to 30% of the meteorologists
Days	It will rain in 30% of the days like tomorrow
Simulations	It will rain according to 30% of the ensemble of tomorrow's forecasts

Participants read each interpretation on a different page and responded to four questions about that interpretation. We randomised the order of presentation of the questions

and the order of presentation of the interpretations within each question and for each participant.

Fluency was assessed with two questions: ease of comprehension and ease of imaginability (.60 < Cronbach's α < .79)[4]. Comprehensibility was measured on a 6-point Likert scale as follows: *"Please rate to what extent each of these forecasts is comprehensible according to you on a scale ranging from -3: Absolutely incomprehensible to +3: Perfectly comprehensible."* There was no 0 middle point in the scale and the answers were recorded for the analyses as ranging from 1 to 6. Imaginability was measured on a 5-point Likert scale as follows: *"Here are five items which relate to the forecast, "There is a 30% chance of rain tomorrow". Please rate to what extent each of these items is easy to imagine on the scale ranging from 1: Not at all to 5: Completely"* We used the average of the ease of comprehension and imaginability as an index of the fluency of each interpretation (min: 1, max: 5.5).

Utility was measured on a 5-point Likert scale using the following question: *"Here are five weather forecasts. Please rate to what extent each of these forecasts is useful. Think, for example, to what extent it would be useful to have that information to choose your clothes or to decide to go to work by bike".* The scale ranged from *1: Not at all useful* to *5: Extremely useful.*

Finally, correctness was assessed on a 6-point scale as follows: *"Imagine that five persons are told, "There is a 30% chance that it will rain tomorrow", and that each person has a different interpretation of the meaning of the forecast. The five interpretations are listed below. Please rate to what extent each of these interpretations is correct according to*

[4] Although it is debated whether Cronbach's alpha is appropriate for 2-item scales because it tends to under-estimate reliability, it remains an accepted and common practice. See here for a discussion on the subject (Eisinga, Grotenhuis, & Pelzer, 2013).

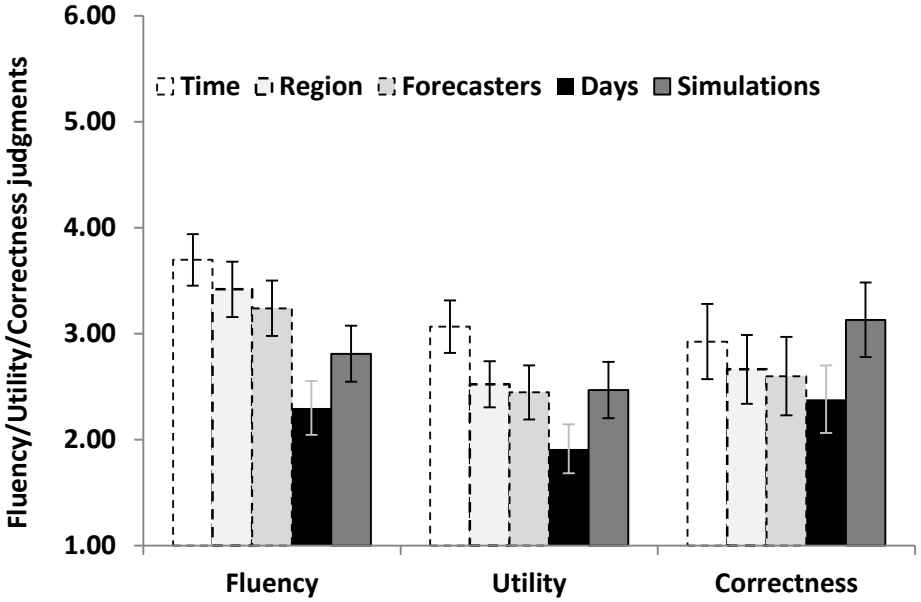
you. Provide your judgment on the scale ranging from -3: Absolutely incorrect to +3: Absolutely correct. There was no 0 middle point in the scale and for the analyses the answers were coded as ranging from 1 to 6.

Finally, participants reported their socio-demographic characteristics.

Results

Fluency, utility and correctness of PoP interpretations

Participants judged the correct Days interpretation to be the least fluent, the least useful and the least correct of the five interpretations (see Figure 1). The most frequent fluency judgments for this interpretation were *absolutely incomprehensible* (34.0%) and *not at all easy to imagine* (43.5%). In contrast, the Simulations interpretation – the correct interpretation devised as a more fluent alternative – was judged to be more fluent, more useful and more correct than the Days interpretation. Yet, the Simulations interpretation was still perceived as less fluent than the three incorrect interpretations (in dotted lines in Figure 1).



Accepted in Journal of Cognitive Psychology (Nov 2018).

This is the version of the article before publication.

Figure 1. Fluency (1-5.5), utility (1-6) and correctness (1-5) perceptions of the Time, Region, Forecasters, Days and Simulations PoP interpretations (N = 92). Bars with dotted lines represent scores for incorrect interpretations, whereas bars with plain lines represent correct interpretations.

<Place Figure 1 about here>

A within-subjects ANOVA showed that fluency perception varied across interpretation, $F(4, 364) = 24.50$, $p < .001$, $\eta_p^2 = .21$. Bonferroni post-hoc tests confirmed the fluency differences. The Days interpretation was less fluent than the Time, Region and Forecasters interpretations, respectively M_{DIFFs}= -1.40, -1.12, -0.94, ps < .001. The Simulations interpretation was more fluent than the Days interpretation, M_{DIFF} = 0.51, p = .006 but was also less fluent than the Time and Region interpretations, M_{DIFF} = -0.89, p < .001; M_{DIFF} = -0.61, p = .026. There was no statistically significant difference in fluency between the Simulations and Forecasters interpretations, M_{DIFF} = -0.43, p = .071.

To test the effect of the fluency of the Days interpretation on correctness perception and to test the mediating role of utility, we conducted a mediation analysis using the SPSS script PROCESS (Preacher & Hayes, 2012). The mediation model is depicted in Figure 2 and shows that fluency predicted the level of correctness of the Days interpretation and that this effect was not mediated by the utility perception of the Days interpretation (the CI interval of the ab path did cross 0).

<Place Figure 2 about here>

Accepted in Journal of Cognitive Psychology (Nov 2018).

This is the version of the article before publication.

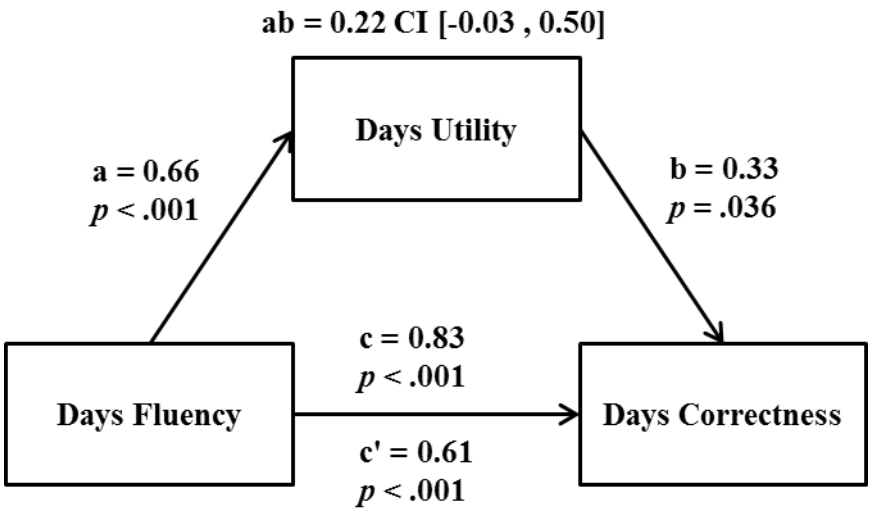


Figure 2. The fluency of the Days interpretations of a PoP forecast predicted its perceived correctness. The effect was not mediated through the utility perception of the Days interpretation (N = 92).

Note: c: Total effect, c': direct effect and ab: indirect effect. *CI*: 95% confidence intervals based on a Monte Carlo simulation on 1,000 samples; the estimates reported are non-standardised estimates.

Testing whether differences in fluency explain differences in correctness

To test whether differences in utility predicted difference in correctness perception and to assess whether this effect was mediated by utility perception, we conducted three mediation analyses on a set of three difference variables. In the first analysis, we compared the Days interpretation to the three incorrect interpretations, in the second, the Simulations interpretation to the three incorrect interpretations, and in the third analysis, we compared the Days interpretation to the Simulations interpretation.

Accepted in Journal of Cognitive Psychology (Nov 2018).
This is the version of the article before publication.

To test whether differences in fluency between the Days interpretation and the three incorrect interpretations accounted for differences in correctness perception, we conducted a mediation analysis on three new difference variables for fluency, correctness and utility. The difference scores were computed by subtracting the average score of the incorrect interpretations from the score of the Days interpretation; hence, lower scores indicated that participants judged the incorrect interpretations to be more fluent, useful or correct than the Days interpretation. For example, DiffDays Fluency = FluencyDays – Mean(FluencyRegion, FluencyTime, FluencyMeteorologists). The results of this mediation analysis are shown in Figure 3, Panel A. The analysis showed that participants judged the incorrect interpretations as being more correct than the Days interpretation because they were more fluent. The effect was partially mediated by utility (as indicated by the fact that the confidence interval of the ab path does not cross 0, while the c' path is still statistically significant).

To test whether differences in fluency between the Simulations interpretation and the three incorrect interpretations accounted for differences in correctness perception, we followed the same procedure except that we computed the difference variables using the Simulations ratings instead of the Days one. For these variables, lower scores indicated that participants judged the incorrect interpretations to be more fluent, useful or correct than the Simulations interpretation. The results of this mediation analysis are shown in Figure 3, Panel B. The analysis showed that participants judged the Simulations interpretation to be more correct than the incorrect ones because the Simulations interpretation was more fluent. The effect was partially mediated by utility.

Finally, to test whether difference in fluency between the Simulations interpretation and the Days interpretation accounted for differences in correctness perception, we conducted a mediation analysis on three new difference variables for fluency, correctness and utility for

which we subtracted responses based on the Simulations interpretation minus responses based on the Days interpretation. For these variables, higher scores indicated that participants judged the Simulations interpretation to be more fluent, useful or correct than the Days one. The results of this mediation analysis are shown in Figure 3, Panel C. The analyses showed that participants judged the Simulations interpretation to be more correct because it was more fluent. The effect was partially mediated by utility. In the three mediation analyses, utility was partly mediating the effect of fluency on correctness.

<Place Figure 3 about here>

Accepted in Journal of Cognitive Psychology (Nov 2018).

This is the version of the article before publication.

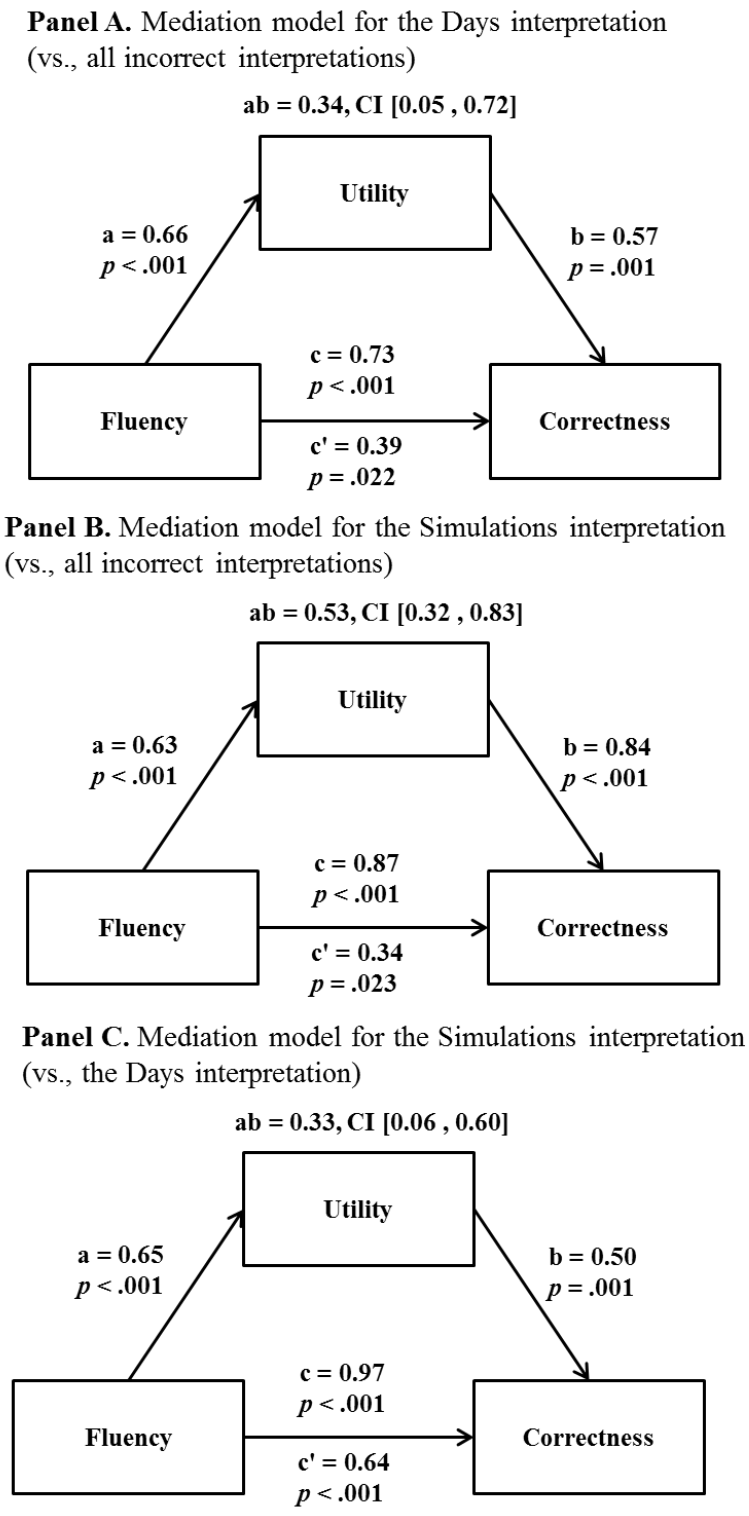


Figure 3. Effect of fluency on perception of correctness with utility as a mediator (N = 92) for three models: the Days (vs. all incorrect interpretations), the Simulations (vs. all incorrect

interpretations) and Simulations (vs. Days interpretation) models, respectively shown in panels A, B and C.

Overall, the results of Study 1 indicated that the correct Days interpretation was perceived to be less correct than the incorrect Time, Region and Forecasters interpretations when it was perceived to be less fluent and less useful. Further, the Simulations interpretation proved to be more fluent, more useful and therefore more correct than the Days interpretation. Although the Simulations interpretation was considered to be less fluent than the three incorrect interpretations, our results provide evidence that it was more fluent than the Days formulation and that hence it was perceived as more correct.

Study 2

The findings of Study 1 have shown that fluency played an important role when assessing whether a PoP interpretation was useful and correct or not. Secondly, an alternative correct interpretation – the Simulations interpretation (i.e., "rain will fall according to 30% of the ensemble of tomorrow's forecasts") was perceived to be more fluent, more useful and, in turn, more correct than the Days interpretation.

Our findings have highlighted a methodological weakness in the design of past research which weakens the reach of its conclusions. The low selection rate of the correct Days interpretation in past research may be partly explained by the fact that this interpretation is less fluent than the other provided in the multiple choice question (Gigerenzer et al., 2005; Juanchich & Sirota, 2015; Morss et al., 2008).

Our findings have indicated that the selection rate of responses is sensitive to probability-irrelevant factors such as the wording of the responses. Hence, to get a better

view of the PoP interpretation it seems that the use of more than one item is required. This is also aligned with the recommendations issued by the American Educational Research Association, the American Psychological Association and the National Council on Measurement in Education (2014).

Study 2 aimed to reassess how participants understand probability of precipitation forecasts in a study featuring a more fluent correct interpretation (i.e., the Simulations interpretation instead of the Days one) and featuring three multiple choice questions tapping into the interpretation of a single PoP forecast. We expected to find that participants would choose the correct interpretation more often than in past research.

Method

Participants. A total of 114 Americans from Amazon Mechanical Turk took part in the survey. A post-hoc sensitivity analysis showed that our sample size allowed us to detect a small effect size of $w = 0.13$ for a chi-squared test with $df = 1$ (both assuming $\alpha = .05$, $1\text{-}\beta = .80$), as planned for a comparison with prior research using the Days interpretation in a similar sample (Juanchich & Sirota, 2016; n = 334).

Table 2 provides an overview of the characteristics of the sample in Study 2 along with the characteristics of the American population. Their ages ranged from 18 to 73 (M = 33.2, SD = 11.5). In the sample, there were slightly more female participants. Most participants had a job, had some experience of higher education and were Caucasian. Fifteen participants did not report their socio-demographics.

Materials and procedure. Participants first read the instruction: *"Imagine that you want to know the weather forecast for the next day in your state. You are given, from a reliable source, the following weather forecast."* We chose a state-wide forecast to allow for the Region interpretation to take place. Participants then received a traditional probability of

precipitation: “There is a 30% chance that it will rain tomorrow”. Along with the PoP, participants read three multiple choice questions assessing their understanding of the forecast (randomly presented). For each they could select one of four choices designed to match the Time, Region and Simulations interpretations along with an “Other” option. The options were presented in a random order to each participant with the Other option always being last and are available in Appendix A.

The first item corresponded to the classic PoP multiple choice question as used in Gigerenzer et al. (2005) except that the Days interpretation was replaced by the Simulations interpretation. The second question featured the probability magnitude corollaries of the PoP interpretations as possible answers: in the Region and Time interpretations people are sure that it will rain; their uncertainty is about where and when. In the third multiple choice question the choices featured ratios instead of normalised frequencies (3 in 10 or 3 out of 10 instead of 30%). According to the American National Weather Service’s definition of PoP forecasts (2015), the Region interpretation is correct when the probability of rain is 100%. To exclude the Region interpretation as potentially correct, we asked participants to report the response that was “absolutely true” (Juanchich & Sirota, 2015).

At the end of the survey, participants provided their socio-demographics. Participants also described whether they lived in a city, how often they sought weather forecasts (from never to every day) and their favourite source of weather forecasts (the internet, TV, phone, radio). The answers to these questions did not impact participants’ PoP interpretations.

The internal reliability of the three questions measuring the PoP interpretations was not satisfactory (Cronbach’s Alpha = .48). However, eliminating the third item led to an increased and satisfactory reliability (Cronbach’s Alpha = .67). The small and not significant correlation between the third question and the other two (.05 and -.11) was taken as a cue that

the question did not address PoP interpretation and therefore we chose not to analyse this data further.

Results

Reassessing how people understand PoP. In the first question, 59% of the participants selected the correct Simulations interpretation, 20% chose the Region interpretation and 10% the Time interpretation, while 12% selected Other and provided a personal interpretation. In the second question, 73% of the participants recognised that the PoP meant that it was uncertain that it would rain at all whereas 24.5% of the participants were certain that it would rain but were uncertain about where (17.5%) or for how long (7%). Only three participants provided a personal response (2.6%). Among the participants who selected the proposed interpretations, the responses in the two interpretation questions were fairly consistent – overall 72% chose the same interpretation to both questions: 2% chose the Time interpretation, 13% opted for the Region interpretation and 57% chose the Simulations interpretation to answer both questions (without the Other option). The Region answer was a more common misinterpretation than the Time one in response to both questions. This rate was possibly inflated by the width of the spatial area covered by the forecast in our context: a state-wide forecast.

For the Region and Simulations interpretations, the participants' answer to the first question made them likely to select the same option in the second question (e.g., 62% and 85%) whereas this was less the case for the Time interpretation (17%).

Among the personal PoP interpretations, most were an exact reformulation of the forecast or a translation of the PoP in a different probabilistic format (e.g. "a 30% possibility", "a 30% likelihood of falling", "the chance of rain is very low", "there is a chance

of rain" and "there is a 3 in 10 chance that it will rain tomorrow"). This rewording tendency is similar to that observed in past work (Gigerenzer et al., 2005; Joslyn et al., 2009; Morss et al., 2008; Murphy et al., 1980; Peachey et al., 2013).

Comparison with the rate of correct interpretation in previous research. We compared the responses given to the first multiple choice question of Study 2 to the results of previous research using the same task in the same population (Amazon Mechanical Turk workers) but with the Days interpretation instead of the Simulations one (Juanchich & Sirota, 2015). The comparison shows that participants selected the correct answer more often than before. Among participants who selected the provided interpretation, when the Simulations interpretation was listed in the possible answers, 67% of the participants selected the correct interpretation, whereas past research using the Days interpretation showed that only 53% of the participants selected the correct interpretation (N = 339; Juanchich & Sirota, 2015). This difference was statistically significant, $\chi^2(1) = 6.79$, $p = 0.009$, Cramer's $V = 0.22$.

General discussion

The findings of Study 1 corroborate the role of metacognitive cues – perceived fluency and utility of the interpretation – in assessing the correctness of the options used to assess people's understanding of probabilistic forecasts in multiple choice questions. Fluency determined the perceived correctness of the probability of precipitation (PoP) interpretations. This effect was partly mediated by the perceived utility of the interpretation. Further, the results of Study 2 offer a more optimistic window on people's abilities to understand probability of precipitation. When provided with a more fluent correct interpretation than the one previously used, a majority – 67% – of the participants selected the correct PoP interpretation. These findings contrast with previously observed rates of correct interpretation

that were generally below 50% (Gigerenzer et al., 2005; Joslyn et al., 2008; Juanchich & Sirota, 2016; Morss et al., 2008; Sink, 1995).

Explaining participants' selection: not just a matter of fluency

Our findings indicated that the Time, Region and Forecasters interpretations were more fluent than the Simulations interpretation (see Figure 1). Yet, participants chose the Simulations interpretation more often than the Time, Region and Forecasters interpretations in the MCQ. This is rather good news as it indicates that fluency is not the only force at work in the selection of a PoP interpretation. The fluency of the Time, Region and Forecasters interpretations could have been driven by experience – people may consider how long and where it will rain more often than the number of simulations in which it rains. This frequency would increase people's familiarity with those interpretations and therefore increase their fluency (Tversky & Kahneman, 1973). This hypothesis is in line with the findings of Lazo et al. (2009) showing that people are mostly interested in where and how long it will rain.

Our findings also highlight the factors that contribute to fluency. Based on past research showing that syntactic (Lowrey, 1998) and lexical cues (Oppenheimer, 2006) contribute to the fluency of a sentence, we could have expected that a longer interpretation would have been less fluent. However, our results indicated that the Simulations interpretation was more fluent than the Days interpretation despite having two extra words.

Reinterpretation of past research

The present findings call for a cautious use of previous data obtained using a single multiple choice question to assess the interpretation of probability of precipitation. We have demonstrated that using a more fluent correct answer leads to a greater number of correct answers, indicating that previous research using non-fluent correct answers (e.g., Gigerenzer et al., 2005; Juanchich & Sirota, 2016; Morss et al., 2008) provided a more pessimistic

picture of PoP understanding than the reality. The possible interpretations presented to people in a multiple choice question PoP should have an equal level of fluency to better reflect how people actually understand PoP forecasts. A stable level of fluency across interpretations is a better setting than the use of multiple choice questions with varying levels of fluency to provide an accurate picture of how people understand PoP. In the present project we have reduced the fluency gap between the correct and incorrect interpretations in the multiple choice questions but we have not closed it. The fluency of the correct Simulations wording was still lower than the Time, Region and Forecasters interpretations, hence, our findings could still provide a negatively biased picture of the number of people who correctly interpret PoP.

An extra point of caution regarding our findings is related to our samples, which we recruited on the web platform Amazon Mechanical Turk. While Amazon Mechanical Turk provides samples more representative of the national population than student samples, they are, however, not a true reflection of the diversity in the global population, on account of the fact that they are all internet users. The composition of the Amazon Mechanical Turk panel of workers varies over the years (Ross, Zaldivar, Irani, & Tomlinson, 2010) but in 2008-2009 the workers were younger, more educated and included a greater proportion of women than the national population (Ross, Zaldivar, Irani, & Tomlinson, 2009). When examining our samples compared to the general US population we noticed similar trends. Our participants differed from the general population in two main aspects: they were younger and more educated. The high level of education in our sample may mean that other, less educated segments of the population would be more likely to misinterpret PoP than was indicated in our findings. Further research focusing on individuals typically not using the internet could complement our observations. In the same vein, our findings could be further developed by

focusing on individuals whose livelihoods depend on the weather, such as in farming or tourism, so as to better appraise the economic consequences of PoP misinterpretation.

Methodological implications

Our findings have offered two methodological implications. Firstly, they demonstrate that processing fluency is an important factor in the design of multiple choice questions. When designing multiple choice questions, researchers should consider the fluency of the provided options because fluency can create a biased picture of people's preferences.

Secondly, our findings highlight the danger of using a single measure within a single method of investigation and indicate that using a variety of approaches may be more suitable to ensure an appropriate measurement. In Study 2 we used three items to tap into participants' interpretations of a PoP, but this did not show a good statistical fit. More items are needed to increase the reliability and validity of a PoP interpretation scale. In the past, Morss et al. (2008) have also called for the use of more questions to assess PoP interpretation. Joslyn et al. (2008) have, for example, provided an interesting example of experiments featuring either several multiple choice questions or several open ended questions to gain some insights into people's PoP interpretations.

Using a wider variety of similarly fluent items brings evidence that the rate of correct PoP interpretations reported in the past was negatively biased by the fact that the correct answer was also perceived to be the least fluent. It should be noted that multiple choice questions may not be the best way to tap into PoP interpretation, but it is unclear which alternative should be favoured. For example, open ended questions (e.g., "what do you think this means?") rarely provide insights into the reference class that people associate with a probability and participants usually simply offer a rewording of the probability (Gigerenzer et al., 2005; Joslyn et al., 2009; Morss et al., 2008; Murphy et al., 1980; Peachey et al., 2013).

Complementing existing methodologies with qualitative methods, such as interviews or focus groups in which the interviewer can guide respondents, could help to clarify the way that people interpret PoP. Further, focusing on the decisions people make could also provide a window into their interpretations. For example, the Time and Region interpretations could be associated with more cautious decisions given that they entail a 100% probability of rain.

By showing that other factors impact how people understand probabilities, we are widening the range of factors to take into account when accounting for variance in probability interpretation. Unlocking the research focus from probability formats should pave the way for a focus on other determinants that could provide useful insights into explaining how people interpret probabilities. For example, a fruitful avenue of investigation would be to focus on the role of differences in the individual's ability to correctly identify the reference class of a probability. Individual differences have been investigated in connection with how much probability people perceived in probability perception tasks (e.g., Reyna, Nelson, Han, & Dieckmann, 2009) or in probability computation tasks (e.g., Bayesian reasoning; Sirota, Juanchich, & Hagmayer, 2014) but scarcely in the interpretation of probabilities. The findings of Gigerenzer and Galesic (2012) indicate, however, that age and numeracy are two significant factors in explaining variation in probability interpretation: the elderly and individuals with lower numeracy skills showed a lower rate of probability interpretation in a medical context.

Practical implications

We recommend that weather forecasters provide the public with a definition of probability of precipitation that is fluent, simple and easy to imagine, while emphasising the fact that precipitation is *not* certain. Using a fluent definition of what PoP means will help weather forecast users to believe in the quality of the interpretation and its utility. Further,

introducing the complementary probability in the PoP (the probability of no rain) should also contribute to an improved interpretation. Our results have provided indirect converging evidence that prompting a reflection on the uncertainty surrounding the event helps people derive the correct interpretation of a PoP. Indeed, the performance of participants in the second MCQ – in which the options made clear what was uncertain (e.g., about the location, the duration or the rainy event itself) – was about 70%. These results are similar to past work that demonstrated in a small sample (n = 33) that providing a PoP with the probability that it will not rain yielded about 64% of correct interpretations (Joslyn et al., 2008). Further research needs to be conducted on a format-related solution, while paying extra attention to the methodological soundness of the materials used.

In the US alone, 235 million people are exposed to almost four weather forecasts every day – almost a billion forecasts altogether. The belief that many people do not interpret PoP correctly may have legitimised its absence from specific sources of weather forecasting. For example, based on informal observations, it is currently rare to see probabilistic forecasts on British or American TV channels. The present findings provide a more positive view on PoP forecasts interpretation, supporting the assumption that weather forecast users can use those forecasts to make more informed and cost efficient decisions. There is still room for improvement though, and educating people on how to correctly interpret probabilities should therefore remain high on the agenda of weather forecast agencies.

Accepted in Journal of Cognitive Psychology (Nov 2018).

This is the version of the article before publication.

Appendix A. Multiple choice questions used in Study 2

Multiple choice question 1.

In your opinion, what is the usual meaning of this forecast?

If the weather conditions are like today, at least a minimum amount of rain...

- ... will fall in 30% of the region tomorrow.
- ... will fall for 30% of the time tomorrow.
- ... will fall according to 30% of the ensemble of tomorrow's forecasts.
- Other, please specify:

Multiple choice question 2.

Based on the forecast, please indicate which of the following predictions is absolutely true:

- It will rain for certain in the state tomorrow, but where in the state is uncertain.
- It will rain for certain in the state tomorrow, but when is uncertain.
- It is not certain it will rain at all in the state tomorrow.
- Other, please specify:

Multiple choice question 3.

Based on the forecast, please indicate which of the following predictions is absolutely true:

- It will rain for 1/3 of the day tomorrow in the state.
- It will rain in 3 cities out of 10 in the state.
- There is a 3 in 10 chance that it will rain on one specific place in the state.
- Other, please specify:

Accepted in Journal of Cognitive Psychology (Nov 2018).

This is the version of the article before publication.

References

Alter, A. L., & Oppenheimer, D. M. (2006). Predicting short-term stock fluctuations by using processing fluency. *Proceedings of the National Academy of Sciences, 103*, 9369-9372.

Alter, A. L., & Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review, 13*, 219-235. doi: http://dx.doi.org/10.1177/1088868309341564

American Meteorological Society. (2008). Enhancing weather information with probability forecasts. Retrieved 2018, from https://www.ametsoc.org/ams/index.cfm/about-ams/ams-statements/statements-of-the-ams-in-force/enhancing-weather-information-with-probability-forecasts/

Borges, B., Goldstein, D. G., Ortmann, A., & Gigerenzer, G. (1999). Simple heuristics that make us smart. In G. Gigerenzer, P. M. Todd & A. B. C. R. G. the (Eds.), (pp. 59–72-59–72): New York: Oxford University Press.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's mechanical turk. *Perspectives on Psychological Science, 6*, 3-5. doi: http://dx.doi.org/10.1177/1745691610393980

De Elia, R., & Laprise, R. (2005). Diversity in interpretations of probability: Implications for weather forecasting. *Americal meteorological Society*, 1129-1143.

Dutton, J. A. (2002). Opportunities and priorities in a new era for weather and climate services. *Bulletin of the American Meteorological Society, 83*, 1303-1311.

Eisinga, R., Grotenhuis, M. t., & Pelzer, B. (2013). The reliability of a two-item scale: Pearson, cronbach, or spearman-brown? *International journal of public health*, 1-6.

Gigerenzer, G., & Galesic, M. (2012). Why do single event probabilities confuse patients? Statements of frequency are better for communicating risk. *British Medical Journal, 344*. doi: 10.1136/bmj.e245

Gigerenzer, G., Hertwig, R., van den Broek, E., Fasolo, B., & Katsikopoulos, K. V. (2005). "A 30% chance of rain tomorrow": How does the public understand probabilistic weather forecasts? *Risk Analysis, 25*, 623-629. doi: http://dx.doi.org/10.1111/j.1539-6924.2005.00608.x

Goodman, J. K., Cryder, C. E., & Cheema, A. (2012). Data collection in a flat world: The strengths and weaknesses of mechanical turk samples. *Journal of Behavioral Decision Making*, 213-224. doi: 10.1002/bdm.1753

Hanrahan, P. O., & Sweeney, C. (2013). Odds on weather: Probabilities and the public. *Weather, 68*, 247-250. doi: 10.1002/wea.2137

Joslyn, S., Limor, N.-G., & Nichols, R. M. (2008). Probability of precipitation assessment and enhancement of end-user understanding. *Bulletin of the American Meteorological Society*, 185-193. doi: http://dx.doi.org/10.1175/2008BAMS2509.1

Joslyn, S., Nadav-Greenberg, L., & Nichols, R. M. (2009). Understandingprobability of precipitation: Assessment and enhancement of end-user understanding. *Bulletin of the American Meteorological Society, 90*, 185-193. doi: http://dx.doi.org/10.1175/2008BAMS2509.1

Juanchich, M., & Sirota, M. (2016). How to improve people's interpretation of probabilities of precipitation. *Journal of Risk Research., 19*, 388-404. doi: 10.1080/13669877.2014.983945

Lazo, J. K., Lawson, M., Larsen, P. H., & Waldman, D. M. (2011). U.S. Economic sensitivity to weather variability. *Bulletin of the American Meteorological Society*. doi: http://dx.doi.org/10.1175/2011BAMS2928.1

Lazo, J. K., Morss, R. E., & Demuth, J. L. (2009). 300 billion served sources, perceptions, uses, and values of weather forecasts. *AMERICAN METEOROLOGICAL SOCIETY, 90*, 785-798. doi: http://dx.doi.org/10.1175/2008BAMS2604.1

Lowrey, T. M. (1998). The effects of syntactic complexity on advertising persuasiveness. *Journal of Consumer Psychology, 7*, 187-206. doi: https://doi.org/10.1207/s15327663jcp0702_04

MetOffice. (2018). The science of 'probability of precipitation'. from https://www.metoffice.gov.uk/news/in-depth/science-behind-probability-of-precipitation

Morgenstern, O., & von Neumann, J. (1943). *Theory of games and economic behavior*: Princeton: Princeton University Press.

Morss, R. E., Demuth, J. L., & Lazo, J. K. (2008). Communicating uncertainty in weather forecasts: A survey of the u.S. Public. *Weather and forecasting, 23*, 974-991. doi: http://dx.doi.org/10.1175/2008WAF2007088.1

Murphy, A. H., Lichtenstein, S., Fischhoff, B., & Winkler, R. L. (1980). Misinterpretations of precipitation probability forecasts. *Bulletin of the American Meteorological Society, 61*, 695-701.

Nadav-Greenberg, L., & Joslyn, S. L. (2009). Uncertainty forecasts improve decision making among nonexperts. *Journal of Cognitive Engineering and Decision Making, 3*, 209-227. doi: http://dx.doi.org/10.1518/155534309X474460

National Weather Service. (2018a). Faq - what is the meaning of pop. from https://www.weather.gov/ffc/pop

National Weather Service. (2018b). Forecast terms, precipitation probability. 2017, from https://www.weather.gov/bgm/forecast_terms

Oppenheimer, D. M. (2006). Consequences of erudite vernacular utilized irrespective of necessity: Problems with using long words needlessly. *Applied Cognitive Psychology, 20*, 139-156. doi: 10.1002/acp.1178

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on amazon mechanical turk. *Judgment and Decision Making, 5*, 411-419.

Peachey, J. A., Schultz, D. M., Morss, R., Roebber, P. J., & Wood, R. (2013). How forecasts expressing uncertainty are perceived by uk students. *Weather, 68*, 176-181. doi: 10.1002/wea.2094

Petrova, P. K., & Cialdini, R. B. (2005). Fluency of consumption imagery and the backfire effects of imagery appeals. *JOURNAL OF CONSUMER RESEARCH, 32*, 442-452. doi: http://dx.doi.org/10.1086/497556

Reber, R., & Schwarz, N. (1999). Effects of perceptual fluency on judgments of truth. *Consciousness and Cognition, 8*, 338-342. doi: http://dx.doi.org/10.1006/ccog.1999.0386

Reyna, V. F., Nelson, W. L., Han, P. K., & Dieckmann, N. F. (2009). How numeracy influences risk comprehension and medical decision making. *Psychological Bulletin, 135*, 943–973-943–973. doi: http://dx.doi.org/10.1136/bmj.38884.663102.AE

Ross, J., Zaldivar, A., Irani, L., & Tomlinson, B. (2009). *Who are the turkers? Worker demographics in amazon mechanical turk.*

Ross, J., Zaldivar, A., Irani, L., & Tomlinson, B. (2010). *Who are the crowdworkers? Shifting demographics in mechanical turk.* Paper presented at the CHI 2010: Imagine all the People, Atlanta, GA, US.

Sink, S. A. (1995). Determining the public's understanding of precipitation forecasts; results of a survey. *National Weather Digest, 19*, 9-15.

Sirota, M., Juanchich, M., & Hagmayer, Y. (2014). Ecological rationality or nested sets? Individual differences in cognitive processing predict bayesian reasoning. *Psychonomic Bulletin & Review, 21*, 198-204. doi: 10.3758/s13423-013-0464-6.

Society, A. M. (2008). Enhancing weather information with probability forecasts. An information statement of the american meteorological society. *Bulletin of the American Meteorological Society, 89*.

Stewart, A. E., Williams, C. A., Phan, M. D., Horst, A. L., Knox, E. D., & Knox, J. A. (2016). Through the eyes of the experts: Meteorologists' perceptions of the probability of precipitation. *Weather and Forecasting, 31*, 5-17. doi: 10.1175/waf-d-15-0058.1

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology, 5*, 207-232. doi: http://dx.doi.org/10.1016/0010-0285(73)90033-9

United States Census Bureau. (2018). American factfinder. 2018, from https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_14_5YR_DP05&src=pt

Zabini, F., Grasso, V., Magno, R., Meneguzzo, F., & Gozzini, B. (2015). Communication and interpretation of regional weather forecasts: A survey of the italian public. *Meteorological Applications, 22*, 495-504. doi: 10.1002/met.1480